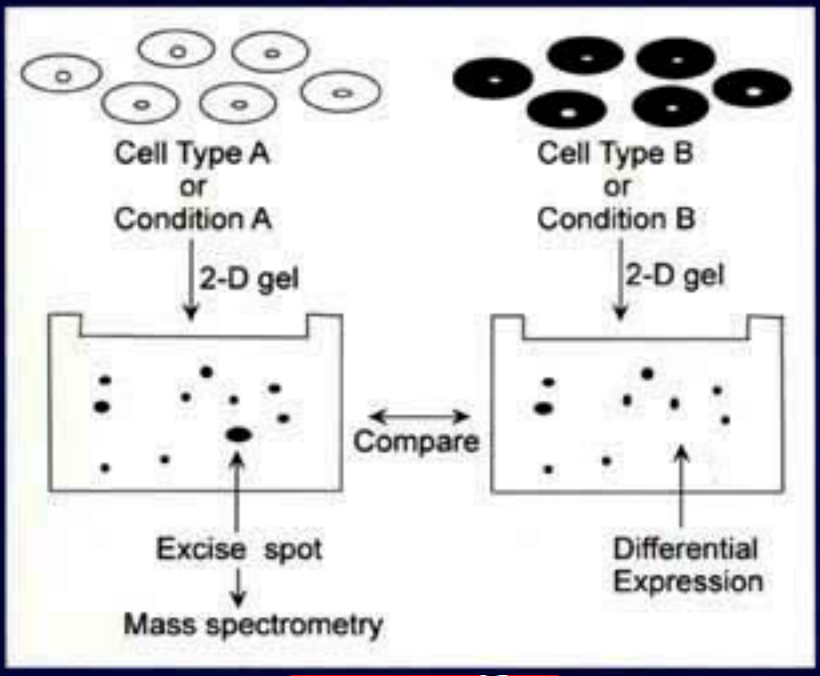
PROTEOMICS

by
Timothy Palzkill



Kluwer Academic Publishers

PROTEOMICS

by

Timothy Palzkill
Baylor College of Medicine

KLUWER ACADEMIC PUBLISHERS
New York / Boston / Dordrecht / London / Moscow

This page intentionally left blank.

PROTEOMICS

eBook ISBN: 0-306-46895-6
Print ISBN: 0-792-37565-3

Created in the United States of America

Visit Kluwer Online at: http://www.kluweronline.com
and Kluwer's eBookstore at: http://www.ebooks.kluweronline.com

CONTENTS

This page intentionally left blank.

PREFACE

Genome sequencing projects have produced an incredible expansion in our knowledge of the DNA sequences of a wide variety of organisms. Thus, the complete set of genes is known for many organisms. However, the function of many of the newly identified genes is not known. Nor is it known how the gene products interact to create a living organism. The field of functional genomics is an attempt to develop large-scale experimental approaches to address these questions. The approach most often associated with functional genomics is microarray hybridization. These experiments provide an assessment of RNA expression levels for all genes simultaneously. Microarrays have become an extremely popular method for understanding transcriptional regulation on a genome-wide scale. However, there is also a great deal to learn at the post-transcriptional level. Large-scale studies of post-transcriptional events are the subject of proteomics.

The name proteomics is traditionally associated with the display of large sets of proteins from a given organism or cell line on two-dimensional (2D) polyacrylamide gels. The ability to associate a spot on a 2D-gel with a known protein is used to create databases of proteins that are expressed in an organism or cell line under defined experimental conditions. This approach is complementary to the generation of databases of mRNA expression levels by microarray hybridization. The combination of technologies permits an assessment of post-transcriptional regulation and post-translation modifications. However, the field of proteomics is rapidly expanding with additional experimental approaches and this book is intended to reflect that expansion. A broader definition of proteomics is used that includes systematic experimental and computational attempts at determining protein-protein interaction maps for an entire organism.

Proteomics is an interdisciplinary science that includes biology, bioinformatics, and protein chemistry. The purpose of this book is to provide the reader with an overview of the types of questions being addressed in proteomics studies and the technologies used to address those questions. The first chapter is a concise outline of the field as it presently stands. The second chapter provides an overview of the use of 2D-gel electrophoresis and mass spectrometry to identify proteins, as well as post-translational

modifications of proteins, on a genome-wide scale. The chapter also includes an assessment of the limitations of this approach and outlines new developments in mass spectrometry that will advance future research. Chapter three describes the use of mass spectrometry to characterize the changes in protein expression profiles in different cell types or in the same cell type under different experimental conditions. The fourth chapter outlines high-throughput recombinant DNA cloning methods used to systematically clone all of the open reading frames of an organism into plasmid vectors for large-scale protein expression and functional studies such as protein-protein interactions with the two-hybrid system.

An important and growing aspect of proteomics is the attempt to generate protein-protein interaction maps for an entire genome. This information is crucial to an understanding of how genes work in concert to generate a working cell. This information, in conjunction with knowledge of transcriptional regulation obtained from microarray experiments, will provide insights into gene function. Chapter five details the experimental approaches used to generate protein-protein interaction maps including the yeast two-hybrid system, mass spectrometry and phage display. Chapter six is a summary of several computational approaches to identify protein interaction networks. Chapter seven describes attempts to create protein microarrays analogous to the DNA chips used to study RNA levels. Protein arrays hold the promise of fast, sensitive protein-protein and protein-ligand interaction mapping on a genome-wide scale. In addition, protein arrays will greatly facilitate drug discovery by allowing the rapid determination of protein targets for a prospective drug. Finally, this chapter covers efforts at determining the function of genes by the activity of the protein products. This involves the large scale cloning, expression and purification of all of the proteins of an organism. This approach has been termed biochemical genomics. Finally, chapter eight describes current limitations and possible future directions for proteomics research.

It is hoped that this book will provide the basis for understanding the field of proteomics. It is not intended to cover every aspect of the field in encyclopedic style but rather to serve as a starting point for more advanced study. Because proteomics is a young and rapidly evolving field, the best approach is to gain a general understanding of the questions and technologies involved and then pursue to the primary literature for detailed information on the latest developments.

Timothy Palzkill

Chapter 1
INTRODUCTION

Whole genome sequence information is now available for many organisms. Sequence analysis of this information reveals many novel genes for which no function can be assigned. Even for well-studied model systems such as *Escherichia coli* and *Saccharomyces cerevisiae,* the specific function of approximately half of the genes is unknown. The challenge of understanding the function of each gene in the genome has led to the development of large-scale, high-throughput experimental techniques that are collectively referred to as functional genomics. These studies include systematic disruption of predicted genes, mRNA expression profiling based on microarray or DNA chip technologies, protein expression profiling using two-dimensional electrophoresis and mass spectrometry, and large scale mapping of protein-protein interactions.

Proteomics is a branch of functional genomics that has arisen in response to the inevitable question posed by the genome sequencing projects, i.e., what are the functions of all the proteins? Proteomics can be defined as the large-scale study of protein properties such as expression levels, post-translational modifications and interactions with other molecules to obtain a global view of cellular processes at the protein level. Because the tools for high-throughput DNA and RNA analysis are not available for protein analysis, the emphasis of functional genomics has been on the mRNA message. However, it is the product of the mRNA, i.e., the protein, which actually carries out the majority of the reactions of the cell. In addition, there is no *a priori* reason to expect that there will be a strict linear relationship between mRNA levels and the protein complement or "proteome" of a cell. Proteomics is therefore a complementary approach to genomics and mRNA expression mapping using microarrays. Finally, most drug targets are proteins; therefore, methods to efficiently analyze the protein complement of cells should contribute directly to drug development.

The activity most often associated with proteomics is fractionating and visualizing large numbers of proteins from cells on two-dimensional

(2D) polyacrylamide gels. These types of experiments have been performed for more than twenty years to build databases of proteins expressed from certain cell or tissue types (Anderson and Anderson, 1996; O'Farrell, 1975). Although this remains an important component of proteomics research, the field has expanded due to the development of additional technologies. Proteomics can be broadly divided into two areas of research: (i) protein expression mapping, and (ii) protein interaction mapping.

Protein expression mapping involves the quantitative study of global changes in protein expression in cells, tissues or body fluids using 2D gel electrophoresis coupled with mass spectrometry. The identity of proteins within spots on 2D gels can be rapidly determined by in-gel proteolysis and peptide mass fingerprinting using mass spectrometry. In addition, recent developments in tandem mass spectrometry using nano-electrospray methods and enabled partial sequence information to be rapidly generated from spots on 2D gels. Thus, it is possible to generate databases of protein expression profiles for various cells and tissues (Rasmussen et al., 1996). In addition, rapid progress has been made in the identification of post-translational modifications of proteins (Oda et al., 2001; Zhou et al., 2001). This information is also being incorporated into protein expression profile databases. The aim of protein expression mapping is to compare the spectrum of proteins expressed in cells or tissues under different environmental conditions or from different disease states. Furthermore, an understanding of post-translational modifications of expressed proteins under different conditions or disease-states is sought. For clinical applications, the objective of protein expression mapping is to identify proteins that are up- or downregulated or modified in a disease-specific manner to use as diagnostic reagents or possible therapeutic targets. For basic research, the goal is to understand how the regulation of protein levels or modifications contributes to the execution and coordination of cellular processes.

Protein-protein interaction mapping involves determining, on a proteome-wide scale, the interaction partners for each of the encoded proteins of a cell or organism. The majority of the proteins within a cell are thought to work in concert with other proteins via direct physical interactions to carry out cellular processes. A great deal can be inferred about the function of a protein through knowledge of its interaction partners. For example, if a protein of unknown function is found to interact with a set of proteins known to be involved in a certain cellular process, the unknown protein can be inferred to contribute to the same process. Therefore, creation of a protein-protein interaction map of the cell would be of great value for

understanding the biology of that cell.

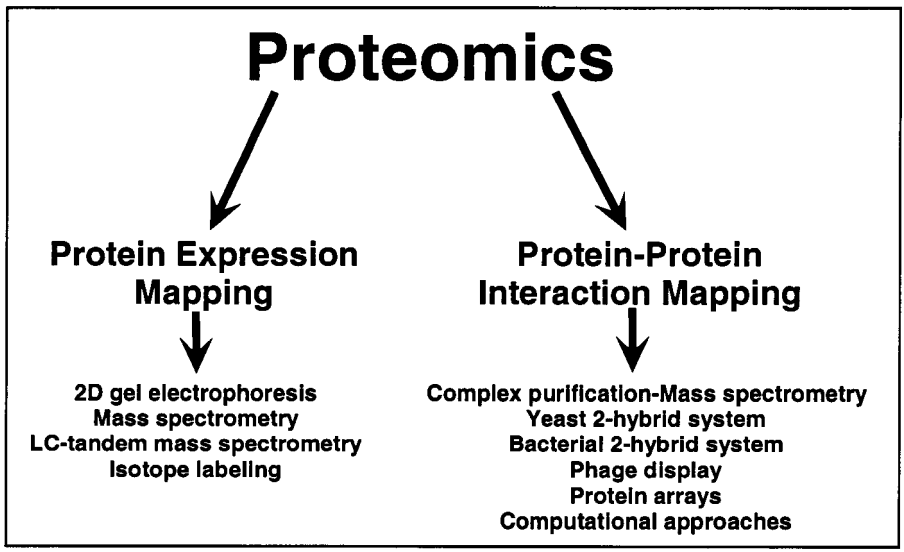


Figure 1.1. Outline of proteomics. Proteomics can be divided into two areas, protein expression mapping and protein-protein interaction mapping. The various experimental approaches used in these areas are listed. LC, liquid chromatography.

A number of technologies are available to study protein-protein interactions. For instance, the yeast two-hybrid system is an *in vivo* method that has been widely used to perform large-scale protein-protein interaction studies (Ito et al., 2001; Uetz et al., 2000). Another widely used approach has been to purify protein complexes from cells and to determine the protein components of the complex by mass spectrometry (Link et al., 1999; Rout et al., 2000). In contrast to the wide application of microarrays to study hybridization of nucleic acids, protein arrays have proved difficult to develop due the fact that proteins possess a precise (and delicate) three-dimensional structure that must be maintained on the surface of the chip. Nevertheless, there have been several reports of large-scale protein arrays and this area is under rapid development. Finally, computational approaches have been developed to predict functional interactions between proteins based on genome sequence data (Eisenberg et al., 2000). These approaches have the advantage that they can be rapidly employed to generate interaction maps for a number of organisms. The result of a computational prediction has been successfully used to guide experimental protein-protein interaction mapping for the rapid generation of a genome-wide interaction map (Newman, 2000).

This page intentionally left blank.

Chapter 2
PROTEIN IDENTIFICATION AND ANALYSIS

One focus of proteomics is to determine the complement of proteins that are expressed in a cell and how this complement changes under different conditions. As such, the ability to accurately identify proteins on a large scale is critical to proteomics studies. Methods for protein characterization, especially mass spectrometry technologies, have greatly improved in accuracy and throughput in recent years. These new technologies have enabled the identification of hundreds to thousands of proteins from organisms and have made the characterization of entire proteomes a realistic goal.

2.1 Protein preparation and separation

Two-dimensional gel electrophoresis

Prokaryotic cells express hundreds to thousands of proteins while higher eukaryotes express thousands to tens of thousands of proteins at any given time. If these proteins are to be individually identified and characterized, they must be efficiently fractionated. One-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) has typically been use to study protein mixtures of $\leq$100 proteins. One-dimensional electrophoresis is useful because nearly all proteins are soluble in SDS, molecules ranging from approximately 10,000 to 300,000 molecular weight can be resolved, and extremely basic or acidic proteins can be visualized. The major disadvantage to one-dimensional gels is that they are not suitable for complex mixtures such as proteins from whole cell lysates.

Two-dimensional separation (2D) involves first separating proteins based on their isoelectric point (pI) using isoelectric focusing (IEF). The isoelectric point is the pH at which there is no net electric charge on a protein. IEF is an electrophoretic technique whereby proteins are separated in a pH gradient. An electric field is applied to the gradient and proteins migrate to the position in the pH gradient equivalent to the pI (Fig. 2.1).

Because the pI of a protein is based on its amino acid sequence, this technique has good resolving power. The resolution can be adjusted further by changing the range of the pH gradient. The use of immobilized pH gradient (IPG) strips has enabled reproducible micropreparative fractionation of protein samples, which is not consistently possible when ampholytes are used in the first dimension (Gorg et al., 2000).

The second step in 2D electrophoresis is to separate proteins based on molecular weight using SDS-PAGE. Individual proteins are then visualized by Coomassie or silver staining techniques or by autoradiography. Because 2D gel electrophoresis separate proteins based on independent physical characteristics, it is a powerful means to resolve complex mixtures proteins (Fig. 2.1). Modem large-gel formats are reproducible and are the most common method for protein separation in proteomic studies.

Limitations of two-dimensional gel electrophoresis

Despite their excellent resolving power, 2D gels are limited in several respects. One problem is the sensitivity and reproducibility of detection. Proteins are expressed in cells over a wide dynamic range of concentrations but the detection range of the Coomassie or silver staining methods is limited. For example, it has been shown by silver staining of 2D gels of protein lysates from *Saccharomyces cerevisiae* that only abundant proteins are identified (Gygi et al., 2000). Even with high sample loads and extended electrophoretic separation, medium to low abundance yeast proteins are not identified. Because these proteins are encoded by approximately 50% of the yeast genes, this observation suggests 2D gel analysis as a technique for proteome characterization is inadequate (Gygi et al., 2000). These results illustrate that, despite the wide use of silver staining, the method has several drawbacks including (i) poor reproducibility, (ii) limited dynamic range, and importantly, (iii) the fact that certain proteins stain poorly or not at all.

The development of fluorescent dyes to visualize proteins from 2D gels may increase the sensitivity and reproducibility of the technique (Steinberg et al., 1996). Staining with dyes such as SYPRO Orange or SYPRO Red is noncovalent and can be performed in a simple one-step procedure after the electrophoretic steps. These dyes bind to the SDS moiety surrounding proteins and therefore show little protein-specific variability (Gorg et al., 2000; Steinberg et al., 1996). Consequently, the dyes provide more uniform staining of proteins and thus reduce protein specific detection artifacts. The detection limit of these fluorescent dyes is in the range of 1-2 nanograms of protein per spot, which is slightly less sensitive than silver staining but, in contrast to silver staining, the linear range of fluorescent staining is over three orders of magnitude. Therefore, staining with

fluorescent dyes holds promise for quantitating the amount of protein in a spot from cells grown under different conditions. Staining proteins with fluorescent dyes, however, does not completely solve the detection problems of 2D gels in that highly expressed proteins can frequently obscure the visualization of proteins expressed at low levels (Gygi et al., 2000).

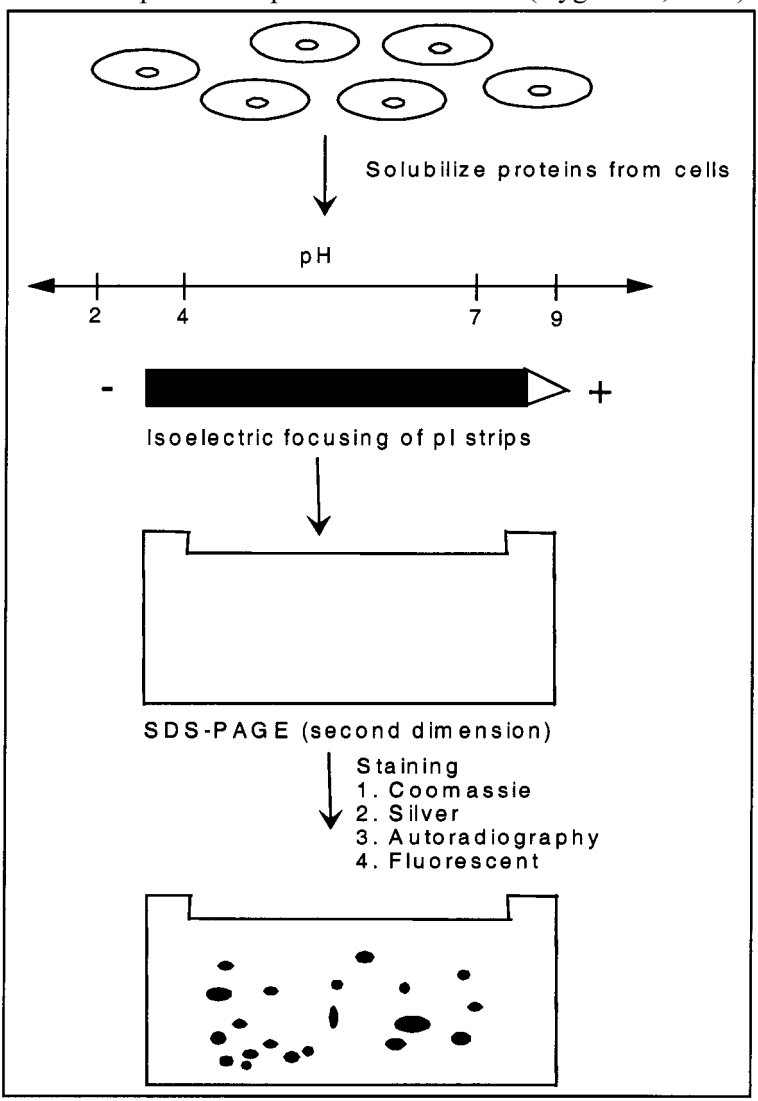


Figure 2.1. Schematic illustration of two-dimensional gel electrophoresis. Proteins are extracted from the organism of interest and solubilized. The first dimension separates proteins based on isoelectric point. The pI strip is reduced and alkylated and applied to an SDS-PAGE gel for separation by molecular weight. Proteins can be visualized using a number of staining techniques.

Another limitation of 2D gels is that membrane proteins are underrepresented. Because membrane proteins account for approximately 30% of total proteins (Wallin and Von Heijne, 1998), this is a serious problem for characterization of the proteome. The relative lack of membrane proteins resolvable on 2D gels can be attributed to thee main factors: (i) they are not abundant, and therefore are difficult to detect by standard staining techniques, (ii) they often possess alkaline pI values, which make them difficult to resolve on the pH gradients most often used for isolelectric focusing, and (iii) the most important reason for under representation may be that membrane proteins are poorly soluble in the aqueous media used for isoelectric focusing (Santoni et al., 2000). Membrane proteins are designed to be soluble in lipid bilayers and are therefore difficult to solubilize in water-based solutions.

If membrane proteins are to be accurately represented, solutions to the thee problems listed above are necessary. New staining techniques such as the fluorescent dyes, as well as methods that allow the loading of milligram quantities of protein onto 2D gels address the problem of abundance (Rabilloud et al., 1994). In addition, proteins with alkaline pI values can be more efficiently separated with new, wide pH range gradients (Gorg et al., 2000). In contrast, problems related to hydrophobicity of membrane proteins are more difficult to solve and progress in this area has been slow. The development of new organic solvents and detergents for the solubilization of membrane proteins is needed.

Another difficulty with 2D gel separations is posttranslational and proteolytic modifications. Although the identification of posttranslational modifications is an important aspect of detailed characterization of the proteome, it can create problems for protein identification. For example, proteolytic degradation of the sample can result in the same protein appearing at several locations on a gel. If this is an abundant protein, it can further obscure low abundance proteins. In addition, the biological significance of proteolytic digestion of a protein is difficult to assess. Posttranslational modifications such as phosphorylation or glycosylation can also place a protein at multiple positions on a gel. However, as described below, rapid progress has been made in the identification of such modifications.

Reducing complexity: Protein fractionation prior to electrophoresis

Because of the difficulties in abundance and compatibility described above, fractionation steps are often performed on protein mixtures prior to 2D gel separation to reduce the complexity of the mixtures. Prefractionation of proteins can be achieved by (i) isolation of cell compartments such as the plasma membrane or organelles such as mitochondria or nuclei, (ii) by

sequential extraction procedures with alternative solubilization capacities such as aqueous buffers versus detergents, or (iii) by fractionation methods such as free flow electrophoresis or chromatography.

The isolation of cell compartments or organelles not only provides a less complex protein mixture for 2D separations, but it also is a means to determine the cellular localization of proteins. Knowledge of cellular location can be an important clue as to the function of a protein. Numerous protocols are available for fractionation of cellular components and the details are dependent on the organism under study. Identification of a protein in a particular compartment can be confirmed by fusing the protein to a readily visible tag such as the green fluorescent protein and determining its cellular location by fluorescence microscopy and imaging (Pandey and Mann, 2000). For example, in a recent study, nuclei from mouse liver cells were isolated and specific nuclear structures named interchromatin granule clusters (IGCs) were purified (Mintz et al., 1999). The purified IGCs were then analyzed by 2D gel electrophoresis and peptide sequencing and mass spectrometry were used to identify the proteins in this structure. Seventeen proteins of unknown function were found among the IGC proteins. The tagging of several proteins with yellow fluorescent protein allowed localization of proteins to the nucleus. Similarly, the proteins of the chloroplast of the garden pea have been catalogued using a related experimental approach (Peltier et al., 2000). By using a combination of methods including chloroplast purification, solubilization of proteins, two-dimensional gel electrophoresis and mass spectrometry, a total of 200 chloroplast proteins were identified in the luminal space and periphery of the chloroplast thylakoid membrane (Peltier et al., 2000). The use of this approach to study the proteomes of organelles and other cellular structures is likely to increase in popularity.

The sequential extraction of protein samples with buffers of increasing solubilizing capacity is another means of fractionating samples. This could involve, for example, an initial extraction with an aqueous buffer followed by an extraction with an organic solvent such as methanol followed by a final extraction with a detergent (Gorg et al., 2000). Such an approach may be useful to fractionate soluble proteins from peripheral membrane proteins and peripheral membrane proteins, in turn, from integral membrane proteins (Santoni et al., 2000). Because membrane proteins and peripheral membrane proteins are poorly soluble in aqueous buffers and may only be partially soluble in organic solvents and detergents, it is important to reduce the complexity of the protein lysate to enrich and concentrate these proteins for subsequent analysis.

A number of affinity-based or chromatography methods have been used to prefractionate protein samples for 2D electrophoresis. For example, proteins of low abundance can be enriched from crude lysates by affinity-

based protein purification strategies, such as the use of an antibody specific to the protein(s) of interest. Another antibody-based approach involves immunoprecipitation of phosphorylated proteins using an anti-phosphotyrosine or anti-phosphoserine antibody (Pandey and Mann, 2000). Alternatively, or in addition to immunoprecipitation, phosphorylated proteins can be affinity purified by immobilized metal affinity chromatography (Ahn and Resing, 2001). After affinity purification, the phosphoproteins can be separated by one or two-dimensional electrophoresis and identified by mass spectrometry as described below. This is an efficient means of identifying post-translational modifications. Other chromatography strategies are less specific, such as the fractionation of proteins by charge or hydrophobicity. The objective in these cases is not to identify highly charged or hydrophobic proteins per se, but rather to provide a reproducible means of reducing the complexity of a whole cell lysate and at the same time concentrating the proteins that are fractionated. By generating a number of 2D protein gel images following a series of general, reproducible chromatographic separations, it may be possible to visualize a large fraction of the proteome of an organism of interest (Fig. 2.2).

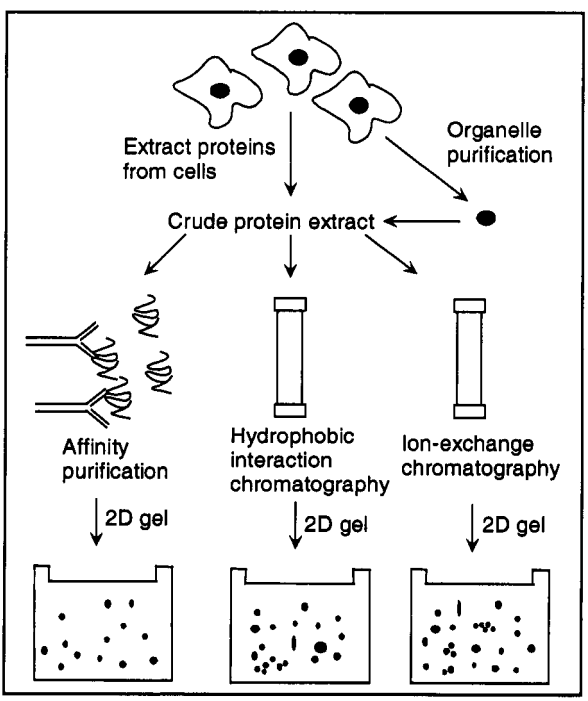


Figure 2.2. Fractionation of protein extracts before 2D gel electrophoresis. Crude lysates can be fractionated by affinity purification or by a number of chromatographic techniques. In addition, organelles or other cellular structures can be purified and lysates from these organelles can be fractionated or separated directly on 2D gels. By repeating this procedure using a number of conditions it may be possible to visualize a large fraction of a cell's proteome.

2.2 Protein identification by mass spectrometry

Basics of mass spectrometry analysis

The ability to visualize spots on a 2D gel, while useful as a fingerprint, is not the same as protein identification. Protein sequencing by the Edman degradation technique is the classical means of determining

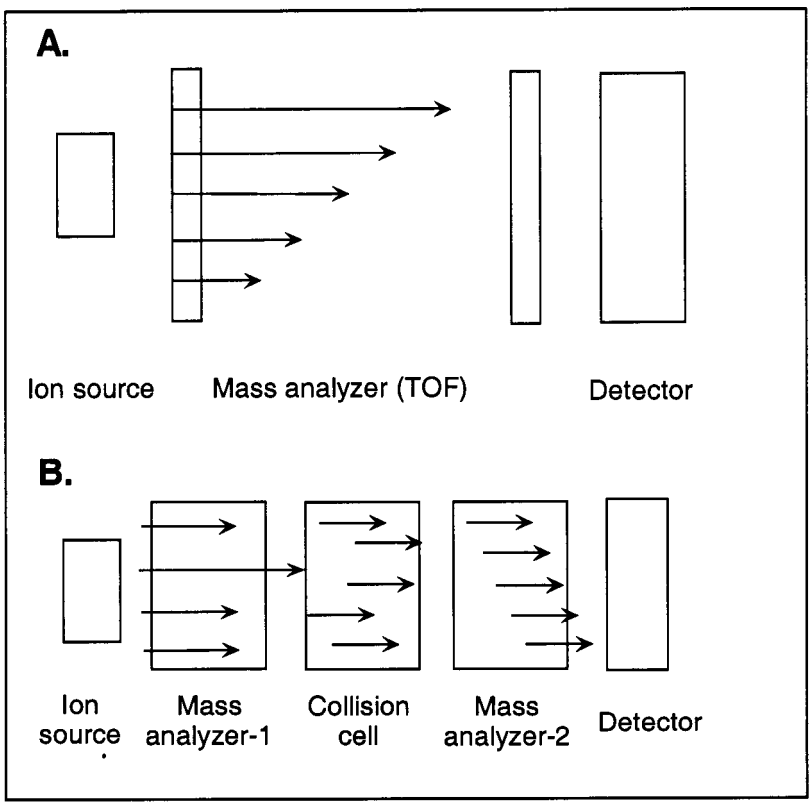


Figure 2.3. A. Mass spectrometer consisting of an ionization source, a mass analyzer and an ion detector. The mass analyzer shown is a time-of -flight (TOF) mass spectrometer. Mass-to-charge (m/z) ratios are determined by measuring the amount of time it takes an ion to reach the detector. B. Tandem mass spectrometer consisting of an ion source, a first mass analyzer, a collision cell, a second mass analyzer and a detector. The first mass analyzer is used to choose a particular peptide ion to send to the collision cell where the peptide is fragmented. The mass of the spectrum of fragments is determined in the second mass analyzer and is diagnostic of the amino acid sequence of the peptide. Figure adapted from Yates III (2000).

protein identity. Because Edman degradation is a low throughput method, however, this results in a bottleneck for determining the identity of spots on 2D gels. One of the major advances in proteomics has been the development of mass spectrometry as a reliable, high-throughput method of protein identification. Mass spectrometry provides extremely sensitive measurements of the mass of molecules and this data can be used to search protein and nucleotide databases directly to identify a protein (Mann, 1996; Wilm et al., 1996). Mass spectrometry relies on the digestion of gel-separated proteins into peptides by a sequence specific protease such as trypsin. Peptides are used rather than proteins because the molecular weight of an entire protein is not sufficiently discriminating for database identification.

Mass spectrometers measure the mass-to-charge ratio (m/z) of ions. They consist of an ionization source that converts molecules into gas-phase ions and a mass analyzer coupled to an ion detector to determine the m/z ratio of the ion (Yates III, 2000). A mass analyzer uses a physical property such as time-of-flight (TOF) to separate ions of a particular m/z value that then strike the detector (Fig. 2.3). The magnitude of the current that is produced at the detector as a function of time is used to determine the m/z value of the ion. While mass spectrometers have been used for many years for chemistry applications, it was the development of reproducible techniques to create ions of large molecules that made the method appropriate for proteomics.

Ionization of biological macromolecules

Matrix-assisted laser desorption ionization (MALDI) creates ions from the energy of a laser with the help of an energy absorbing matrix (Cotter, 1999). The molecules to be ionized are desiccated in a crystalline matrix and the laser causes excitation of the matrix and the ejection of ions into the gas-phase. This method of ionization is often used in conjunction with time-of-flight detection (MALDI-TOF) to identify proteins by peptide mass fingerprinting (Fig. 2.4) (Henzel et al., 1993). The masses of peptides derived from an in-gel proteolytic digestion of protein spots from a 2D gel are measured and searched against a computer generated list of peptides created by a simulated digestion of a protein database using a specific protease such as trypsin. The accuracy of the mass measurement is often sufficient to identify proteins from completely sequenced genomes, such as the bacterium *Haemophilus influenzae* (Langen et al., 2000). This is possible because the masses of all of the tryptic peptides from the predicted open reading frames can be precisely calculated for comparison to the mass data. The power of this approach has increased due to advances in automation such that hundreds of proteins can be visualized, excised, digested enzymatically, and their mass determined and automatically searched against databases (Berndt et al., 1999).

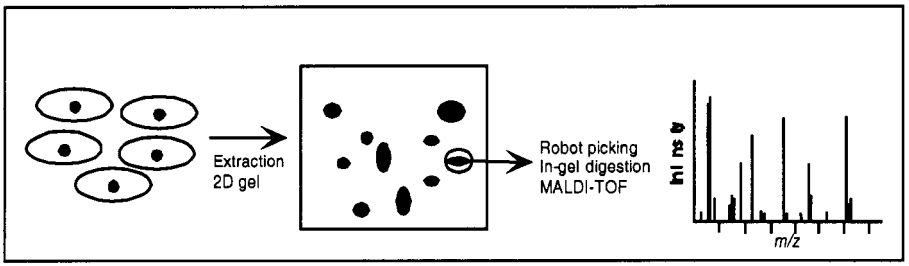


Figure 2.4. Peptide fingerprinting by MALDI-TOF mass Spectrometry. Proteins are extracted and separated on by 2D gel electrophoresis. A spot of interest is excised from the gel, digested with trypsin, and ionized by MALDI. The precise mass of proteolytic fragments is determined by time-of-flight mass spectrometry. The identity of the protein is determined by comparing the peptide masses with a list of peptide masses generated by a simulated digestion of all of the open reading frames of the organism.

Electrospray ionization (ESI) creates ions by holding a liquid at a high potential difference. This results in a local separation of charges. The repulsion of these charges overcomes the surface tension of the liquid and gives rise to a spray of charged droplets of solvent containing analyte. Cycles of evaporation remove the solvent and result in the formation of ions. The ions enter a mass analyzer such as TOF and give rise to an *m/z* spectrum (Miranker, 2000; Yates III, 2000). A variation of the method is nano-electrospray ionization, which is commonly used in proteomic studies (Wilm et al., 1996; Miranker, 2000). This method involves the use of a miniaturized electrospray source consisting of a metal-coated glass capillary with an inner diameter of 1 μM. The tip of the capillary is held at a potential difference of 1-2 kV with respect to the orifice of the mass analyzer, which results in spray droplets that are approximately 100 times smaller in volume than those produced by conventional electrospray sources (Wilm et al., 1996). An advantage of nano-electrospray ionization is that little of the sample is lost in large droplets from which biomolecules cannot be ionized. In addition, very small amounts of a sample can be subjected to mass spectrometric analysis over a long period of time. This greatly facilitates the use of tandem mass spectrometry as described below. Finally, using nanoelectrospray, it is possible to ionize biomolecules from aqueous buffers at neutral pH and room temperature. This property has lead to new applications for mass spectrometry in the study of protein complexes (Miranker, 2000).

Tandem mass spectrometry

Tandem mass spectrometry (MS/MS) is another common approach used for protein identification. In this method, proteins are digested and the resulting peptides are ionized directly from the liquid phase by

nanoelectrospray ionization. The peptide ions are then sprayed into a tandem mass spectrometer (Fig. 2.3B). This instrument consists of the ion source, a mass analyzer, a gas-phase collision cell, a second mass analyzer and an ion detector. The first mass analyzer is used to resolve the peptides in the mixture and isolate one species at a time that is then sent to the collision cell. The peptides are fragmented by collisions with an inert gas molecule such as argon. For peptide ions, fragmentation occurs at or around the amide bond to create a ladder of fragments that are diagnostic of the structure of the peptide. The mass of the fragments is precisely determined in the second mass analyzer to yield amino acid sequence information on the peptide (Fig. 2.5). The partial amino acid sequence determined from several of the peptides, called peptide sequence tags, can be used to search protein databases to identify the protein of interest (Mann, 1996; Wilm et al., 1996). The peptide sequence tags contain a short stretch of sequence from within a peptide as well as the mass, from the fragmentation point, to the amino and carboxyl termini of the peptide (Wilm et al., 1996). This information can be used to search nucleic acid databases in addition to the protein databases. This is done using computer programs that convert the short amino acid sequence information into a degenerate nucleotide sequence pattern that can be used with a string searching algorithm to find matches within nucleotide databases (Mann, 1996).

The major advantage of the tandem mass spectrometry approach compared to MALDI peptide fingerprinting, is that the sequence information obtained from the peptides is more specific for the identification of a protein than simply determining the mass of the peptides. This permits a search of expressed sequence tag nucleotide databases to discover new human genes based upon identification of the protein. This is a useful approach because, by definition, the genes identified actually express a protein.

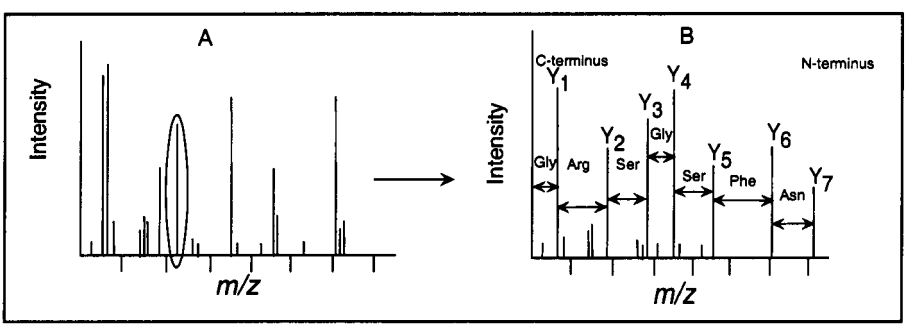


Figure 2.5. Tandem mass spectrometry. **A.** A peptide mixture is electrosprayed into the mass spectrometer. Individual peptides from the mixture are isolated (circled peptide) and fragmented. **B.** The fragments from the peptide are mass analyzed to obtain sequence information. The fragments obtained are derived from the N or C terminus of the peptide and are designated "b" or "y" ions, respectively. The spectrum shown indicates peptides that differ in size by the amino acids shown.

Multidimensional liquid chromatography and tandem mass spectrometry

Current proteomics studies rely almost exclusively on 2D gel electrophoresis to resolve proteins before MALDI-TOF or ESI-MS/MS approaches. A drawback of the 2D gel approach is that it is relatively slow and work intensive. In addition, the in-gel proteolytic digestion of spots followed by mass spectrometry is a one-at-a-time method that is not well suited for high throughput studies. Therefore, considerable effort is being directed towards alternate methods for higher throughput protein characterization.

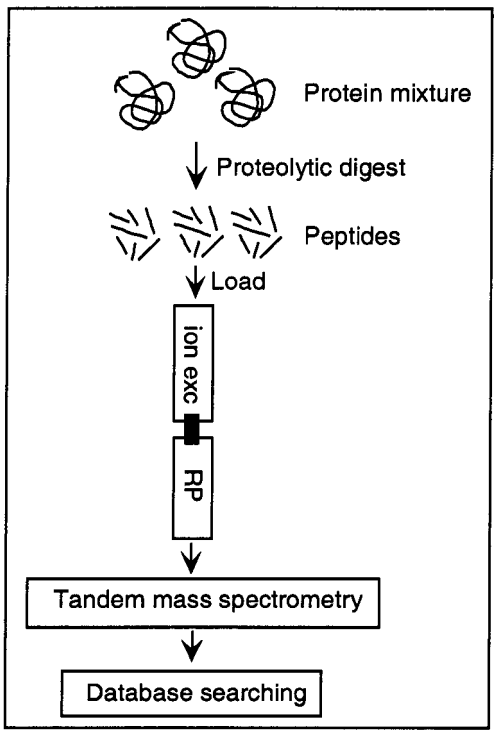


Figure 2.6. LC-tandem mass spectrometry to examine complex mixtures. The mixture of many different proteins is digested to yield peptides and the peptides are resolved into fractions by cation exchange chromatography followed by reverse phase chromatography. The fractionation steps resolve the peptides into fractions that be processed by tandem mass spectrometry to yield sequence information suitable for database searching.

One attempt to overcome these disadvantages has been to use multidimensional liquid chromatography (LC) followed directly by tandem mass spectrometry to separate, fragment and identify proteins (Link et al., 1999). In this process, a denatured and reduced protein mixture is digested with a protease to create a collection of peptides (Fig. 2.6). The peptide mixture is applied to a cation exchange column and a fraction of these peptides are eluted based on charge onto a reverse-phase column. The

peptides are then washed, and eluted from the reverse-phase column based on hydrophobicity into a tandem mass spectrometer to obtain sequence information (Link et al., 1999). The sequence information is used to search genomics databases to identify the proteins in the mixture (Fig. 2.6). The method has been used to characterize the components of the *Saccharomyces cerevisiae* ribosome. For these experiments, ribosomal particles were isolated and the component proteins were trypsinized and fractioned as described above. A total 95 unique proteins were identified from the initial ribosomal particle mixture (Link et al., 1999). 75 of the 78 proteins previously reported to be part of the ribosome were contained among these 95 proteins. Thus, the method was highly effective even without 2D gel separation of proteins. The same experiment was also performed using a total yeast extract rather than purified ribosomal particles. These experiments identified 189 unique proteins among which were 71 of the 78 predicted ribosomal proteins (Link et al., 1999). Therefore, this technique exhibits excellent resolving power. The method has the additional advantage of high throughput because of the lack of 2D gels. Ultimately, the approach's usefulness will depend on the number of proteins and the dynamic range of concentrations of proteins that can be identified using the technique.

Fourier transform ion cyclotron resonance mass spectrometry

Another alternative to 2D gels is the combination of capillary isoelectric focusing (CIEF) and Fourier transform ion cyclotron resonance (FTICR) mass spectrometry (Jensen et al., 1999). The advantages of CIEF are speed, efficiency and the potential for automation of capillary electrophoresis. In addition, protein separations based on pI differences as small as 0.013 pH units have been achieved using the technique (Jensen et al., 1999). After CIEF focusing, proteins are detected by electrospray ionization and FTICR mass spectrometry. FTICR provides extremely high mass resolution, sensitivity and accuracy. Because of the accuracy of the mass determination, proteolytic digestion of proteins before mass spectrometry is not required and hundreds of proteins can be identified in a single run (Jensen et al., 1999).

The use of CIEF in combination with FTICR has been demonstrated in an analysis of the *E. coli* proteome (Jensen et al., 1999). For these experiments, *E. coli* was grown in a medium depleted of rare isotopes in order to increase the mass measurement accuracy. The high abundance isotopes are present at approximately 98.89% ^{12}C, 99.63% ^{14}N and 99.985% ^{1}H. For peptides, the presence of rare isotopes does not significantly change the spectra but with undigested proteins, mass accuracy can be limited by the broadened distribution of ions of any given protein due to the incorporation

of rare isotopes such as ^{13}C and ^{15}N (Jensen et al., 1999). The use of isotope-depletion also improves the signal-to-noise ratio by concentrating the signal from a protein into a single peak.

A single experiment using isotope-depletion followed by CIEF-FTICR generated mass information on hundreds of proteins (Jensen et al., 1999). For example, when data from such an experiment was visualized in a format similar to a 2D gel with putative protein masses plotted versus the scan number, i.e., pI, a virtual 2D display containing approximately 900 spots was generated (Jensen et al., 1999). The identity of the spots could often be determined directly by mass because of the high accuracy of FTICR. Thus, the method has the potential to yield information similar to that obtained from 2D gels but with much less effort. One potential problem is that the pI and molecular mass measurement are insufficient for protein identification in some cases. Incorporating tandem mass spectrometry into the technique has addressed this problem. It has been shown that a peak identified by FTICR can be trapped, fragmented in a collision cell, and mass analyzed to determine the mass of fragments (Jensen et al., 1999). The partial sequence information obtained from the fragments provided unambiguous protein identification (Jensen et al., 1999). Further improvements in this on-line fragmentation method may eventually make it possible to analyze an entire proteome in a single experiment. In addition, combination of the FTICR mass analysis method with other means of protein fractionation may allow identification of proteins of low abundance or low solubility.

2.3 Identification of post-translational modifications

Overview

Post-translational modification of proteins plays a critical role in cellular function. For, example protein phosphorylation events control the majority of the signal transduction pathways in eukaryotic cells. Therefore, an important goal of proteomics is the identification of post-translational modifications. Proteins can undergo a wide range of post-translational modifications such as phosphorylation, glycosylation, sulphonation, palmitoylation and ADP-ribosylation. These modifications can play an essential role in the function of the protein and mass spectrometry has been used to characterize such modifications.

An example of the general approach is provided by a study of post-translational modification of the receptor for the peptide hormone, endothelin (Roos et al., 1998). Endothelin, a 21-amino acid peptide, is the strongest vasoconstrictor known; it elicits physiological effects on cellular

development, vasoconstriction, and mitogenesis (Rubanyi and Polokoff, 1994). The endothelin receptor B is a member of the G-protein-coupled receptor superfamily. It contains seven transmembrane-spanning regions and is post-translationally modified by glycosylation, phosphorylation and palmitoylation (Rubanyi and Polokoff, 1994). Endothelin receptor B was affinity purified to near homogeneity using immobilized endothelin peptide and was digested into peptides using trypsin. The peptides were subjected to fingerprinting by MALDI-TOF mass spectrometry as shown above in Fig. 2.4. By calculating the expected fragment sizes with and without a phosphate (80 Daltons), it was determined that the receptor molecule was phosphorylated and the position was localized to specific peptides (Roos et al., 1998). A similar analysis indicated the presence of a palmitoylated (238 Daltons) peptide. The precise position of phosphorylation and palmitoylation was then determined using collision-induced peptide fragmentation and tandem mass spectrometry (Fig. 2.5) (Roos et al., 1998).

Identification of phosphorylated proteins

The endothelin B receptor is an example of characterization of a homogeneous, affinity purified protein (Roos et al., 1998). Significant progress has been made in the development of techniques for more high-throughput identification of phosphorlyation events. Analysis of large sets of phosphorylated proteins is facilitated by the availability of affinity purification methods such as anti-phosphotyrosine or anti-phosphoserine antibodies or metal affinity chromatography (Neubauer and Mann, 1999; Soskic et al., 1999). These methods are not specific to a particular protein but rather are used to fractionate all proteins that are phosphorylated.

The power of the anti-phosphotyrosine or anti-phosphoserine antibodies is illustrated by a study of phosphorylated proteins in mouse fibroblasts (Soskic et al., 1999). In this study, crude lysates of proteins from fibroblasts that had been either treated or not treated with plateletderived growth factor (PGDF) were fractioned by 2D gel electrophoresis and 260 phosphoproteins were detected with anti-phosphotyrosine antibody while 300 were detected with anti-phosphoserine antibody (Soskic et al., 1999). The identified proteins were then in-gel digested with trypsin and the resulting peptides were analyzed by MALDI-TOF mass spectrometry. The identity of proteins was confirmed and the precise position of phosphorylation was determined by the use of tandem mass spectrometry (Soskic et al., 1999). This effort not only identified proteins that appear to be phosphorylated in response to PDGF treatment, but also examined the kinetics of phosphorylation by treating the cells with PDGF for various times before resolving the proteins by 2D gel electrophoresis (Soskic et al., 1999). A complete map of the positions of particular proteins on the 2D gels

will provide a very powerful tool because it will be possible to study the phosphorylation and expression patterns of the proteins using 2D gel electrophoresis in the absence of mass spectrometry. This will be useful for the analysis of the complex phosphorylation events that occur during signal transduction where it is desirable to perform multiple experiments under many different conditions.

Another means of moving beyond pure protein preparations to high-throughput characterization of proteomes is to enrich for phosphopeptides from complex mixtures by metal affinity chromatography (Andersson and Porath, 1986). Using this method, protein mixtures are proteolyzed to create peptides and phosphorylated peptides are enriched by metal affinity chromatography and subsequently identified by mass spectrometry. This method is limited, however, because in many cases phosphopeptides absorb poorly or nonphosphorylated peptides absorb nonspecifically to the metal affinity resins (Ahn and Resing, 2001).

Two recent approaches may improve the high-throughput identification of phosphoproteins. These techniques employ an enrichment achieved via the selective chemical modification of phosphoproteins within complex protein mixtures (Oda et al., 2001; Zhou et al., 2001). In one of the approaches, the protein mixture is first proteolytically digested and the peptides are reduced and alkylated to eliminate reactivity from cysteine residues (Zhou et al., 2001). After the chemical protection of the N- and C-termini, phosphorylated residues are converted to phophoramidate adducts by carbodiimide condensation. This reaction creates free sulfhydryl groups in place of the phosphates and the sulfhydryls are then used to capture the peptides onto glass beads. The beads are washed and the peptides are eluted with trifluoroacetic acid to regenerate the phosphopeptides for analysis by mass spectrometry (Fig. 2.7A) (Zhou et al., 2001). It was demonstrated with digests of the phosphorylated test proteins β-casein and myelin basic protein that both phosphoserine and phosphotyrosine containing peptides can be enriched and detected with the method. When the method was used on a whole cell protein lysate from *Saccharomyces cerevisiae,* phosphate-containing peptides derived from a number of proteins were detected (Zhou et al., 2001). All of these peptides contained either phosphoserine or phosphothreonine. It was suggested that phosphotyrosine containing peptides were not detected because of their low abundance. This is consistent with the additional finding that all of the phosphopeptides identified were derived from high abundance proteins such as glycolytic enzymes (Zhou et al., 2001). Therefore, this method will require other fractionation procedures to identify low abundance phosphoproteins. Nevertheless, the technique represents a significant improvement in detection of phosphoproteins and should allow for rapid characterization of these proteins from proteomes.

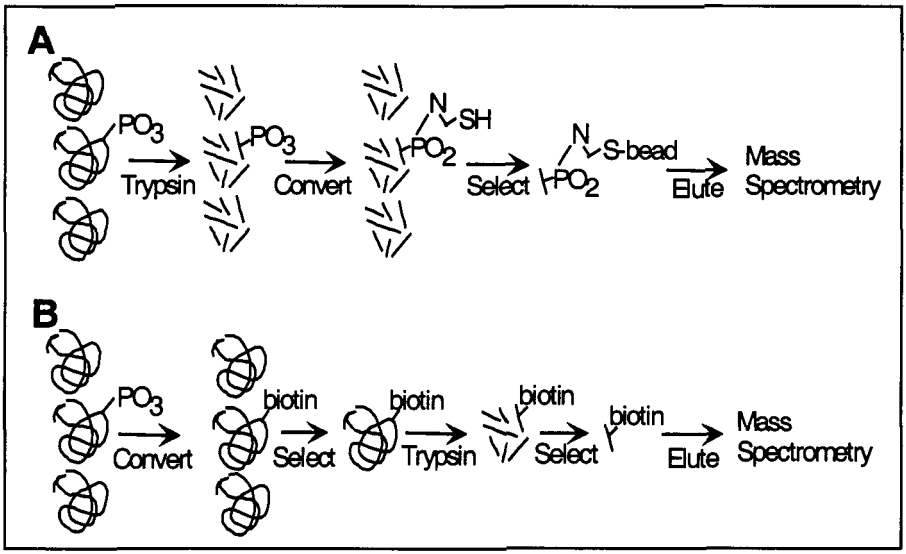


Figure 2.7. Identification of phosphoproteins by site-specific chemical modification. **A.** Method of Zhou et al. (2001) involves trypsin digest of complex protein mixture followed by addition of sulfhydryl groups specifically to phosphopeptides. The sulfhydryl group allows capture of the peptide on a bead. Elution of the peptides restores the phosphate and the resulting phosphopeptide is analyzed by tandem mass spectrometry. **B.** Method of creates a biotin tag in place of the phosphate group. The biotin tag is used for subsequent affinity purification. The purified proteins are proteolyzed and identified by mass spectrometry.

The second method also relies on site-specific chemical modification of phosphoproteins (Oda et al., 2001). It involves the chemical replacement of phosphates on serine and threonine residues with a biotin affinity tag (Fig. 2.7B). The replacement reaction takes advantage of the fact that the phosphate moiety on phosphoserine and phosphothreonine undergoes β-elimination under alkaline conditions to form a group that reacts with nucleophiles such as ethanedithiol. The resulting free sulfydryls can then be coupled to biotin to create the affinity tag (Oda et al., 2001). The biotin tag is used to purify the proteins subsequent to proteolytic digestion. The biotinylated peptides are isolated by an additional affinity purification step and are then analyzed by mass spectrometry (Oda et al., 2001). This method was also tested with phosphorylated β-casein and shown to efficiently enrich phosphopeptides. In addition, the method was used on a crude protein lysate from yeast and phosphorylated ovalbumin was detected. Thus, as with the method of Zhou et al. (2001), additional fractionation steps will be required to detect low abundance phosphoproteins.

The site-specific chemical modification of post-translation

modifications followed by affinity chromatography and mass spectrometry holds promise for the high throughput detection of these modifications. Ultimately, it would be useful to quantitatively detect post-translation modifications of proteins from cells grown under different conditions or from different cell types. One important application would be to detect alterations in post-translation modifications in different disease states. For example, it would be of great interest to examine phosphorlyation patterns in normal human tissue as well as tumors. Such information would be useful in understanding disease progression and well as providing markers for diagnosis. With the rapid changes in technology taking place, these goals are likely to be realized in the near future.

This page intentionally left blank.

Chapter 3
PROTEIN EXPRESSION MAPPING

Nearly all cellular functions are determined by the activity of proteins. Proteins act as catalysts, receptors or structural components that are required for the life of a cell. Many cellular processes are performed by complexes of several different proteins. It is essential that the protein components of these complexes be expressed at the same time and in the same place for the cell to function efficiently. Therefore, an understanding of cellular function at the molecular level requires knowledge of the patterns of expression of all of the component proteins.

Protein expression mapping, as it is commonly implemented, is the quantitative study of global changes in protein expression in tissues, cells or body fluids using two-dimensional gels and mass spectrometry. The objective of these studies is to identify proteins that are up- or down-regulated in response to an environmental stimulus or in a disease specific manner. In this sense, protein expression mapping is analogous to array-based mRNA expression profiling based on hybridization of labeled RNA probes to cDNAs or oligonucleotides that have been immobilized on chips. Protein expression mapping has an advantage over monitoring mRNA levels in that it is a direct measure of the protein product of a gene. However, as described below, several technical challenges must be overcome before protein expression mapping achieves the ease of use of mRNA expression profiling.

3.1 Protein expression mapping by 2D gel electrophoresis and mass spectrometry

Overview

The use of 2D gel electrophoresis and mass spectrometry to identify proteins was discussed in Chapter 2. Protein expression mapping involves the use of these methodologies to compare expression patterns in different cell types or in the same cell type that has been exposed to different

environmental conditions (Fig. 3.1). The most direct route to understanding how and why these experiments are done is to outline some specific examples.

Protein expression mapping in mammalian systems

The utility of protein expression mapping using 2D gel electrophoresis and mass spectrometry has been demonstrated for several experimental systems. One application has been to assess the differences in protein expression between normal and cancerous cells. For example, expression mapping has been used to identify protein markers for bladder cancer (Ostergaard et al., 1999). This was accomplished by identifying proteins released into the urine of patients with and without bladder cancer using 2D electrophoresis and mass spectrometry.

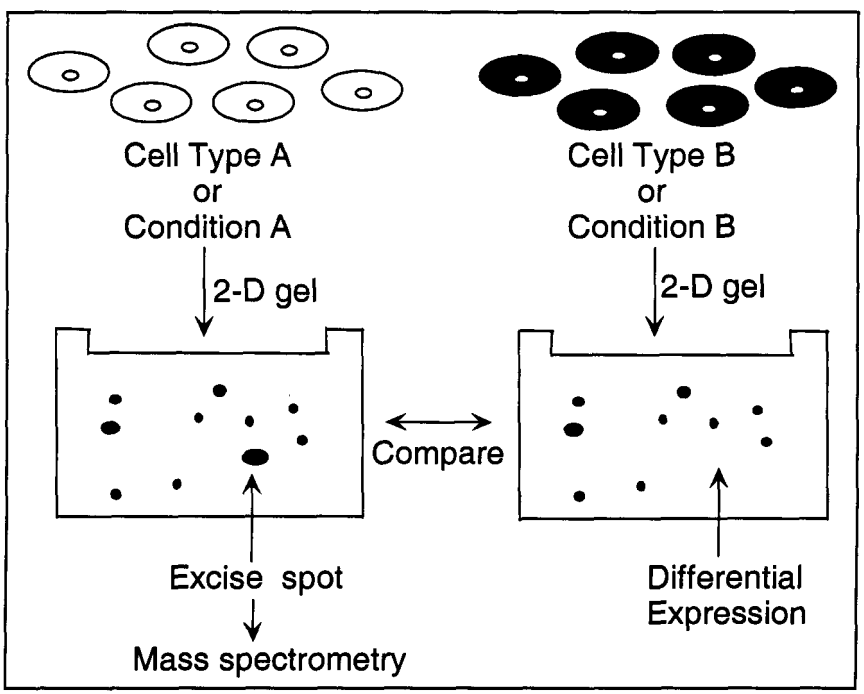


Figure 3.1. Protein expression mapping using 2-D electrophoresis and mass spectrometry. The purpose is to compare protein expression patterns between cell types or in the same cell type under different growth conditions. Proteins are extracted from the different cell types and separated by 2D gel electrophoresis. Image analysis programs are used to compare the spot intensities between gels and identify proteins that are differentially expressed. The protein of interest is excised from the gel and its identity is determined by mass spectrometry. The power of the method increases greatly if the identity of a large number of proteins on the gel is known and present in a database because information can then be obtained without further mass spectrometry.

Squamous cell carcinoma is a type of bladder cancer that is highly malignant and thus the success of therapy is dependent on early detection. Consequently, the identification protein markers that are diagnostic of squamous cell carcinomas would be very advantageous. Protein expression mapping depends on the development of an extensive database of protein expression patterns in tissue and cell types. For the bladder cancer study, this involved comparing the expression profiles between cells from squamous cell carcinomas and normal urothelium cells (Rasmussen et al., 1996). This approach requires knowledge of the protein composition of normal urothelial cells, the tumor cells, and the urine itself. The bladder cancer experiments were facilitated by the availability of a large database of protein expression profiles from keratinocytes. Squamous cell carcinomas are composed of a cell type that is similar to keratinocytes both in terms of morphology and protein expression profile. It was therefore possible to use the information in the keratinocyte database to identify proteins externalized to the urine that are specific to squamous cell carcinomas (Ostergaard et al., 1999). One such protein is a low-molecular weight calcium binding protein named psoriasin. Based on the expression pattern of psoriasin obtained from the comparison of normal and cancer cells by 2-D gels, antibodies were raised against psoriasin and used to show that it is found in the urine of patients with bladder cancer but not in normal patient controls (Ostergaard et al., 1999). Therefore, psoriasin is a candidate marker to be used alone or in combination with other markers as a diagnostic agent for bladder cancer.

Protein expression mapping by 2D gels and mass spectrometry has also been used to identify markers specific for breast cancer (Page et al., 1999). In this study, the protein expression profiles of normal adult human luminal and myoepithelial breast cells were compared. The cell types were purified from breast tissue in sufficient quantities for proteomic studies by using an immunoaffinity cell sorting technique (Page et al., 1999). After cell lysis and protein separation by 2D electrophoresis, the proteins were detected using a fluorescent dye that binds noncovalently to the SDS moiety that attaches to proteins during SDS-PAGE. The stained gels were scanned and a computer program was used to analyze each spot between the sets of luminal or myoepithelial protein expression profiles (Page et al., 1999). A total of 170 proteins that exhibited at least two-fold changes in expression between the cell types were identified. Fifty-one of these proteins were identified by tandem mass spectrometry. These included several enzymes of muscle-specific origin that were expressed at higher levels in myoepithelial cells. In addition, the annexin II protein, which is differentially regulated between normal and malignant breast epithelial cells, was also identified. The majority of the proteins identified, however, were abundant cytoskeleton proteins such as keratin (Page et al., 1999). Whether these proteins will serve as useful diagnostic markers or drug targets awaits further study.

Protein expression mapping has also been used to study cellular differentiation. A model system used for studying differentiation of the mammary gland involves a well defined cellular modification in a rat mammary adenocarinoma cell line that results in the formation of hemispheric structures called ìdomesî (Zucchi et al., 2001). After exposure to a differentiating agent such as DMSO, the rat adenocarcinoma cell line LA7, but not the 106A10 cell line, is able to form domes. A proteomic approach was undertaken to identify differentially expressed proteins in the LA7 cells under the influence of DMSO versus 106A10 cells (Zucchi et al., 2001). Total cell extracts of the cell lines were run on 2D gels and approximately 2000 proteins were separated per gel. Image analysis was performed using a computer program to identify more than 200 differentially expressed proteins (Zucchi et al., 2001). Fifty of these proteins were identified by excision of the spot from the gel followed by MALDI/TOF mass spectrometry. Two of the identified proteins were studied further because of their striking expression patterns. Tm-5b protein was expressed 50-60 times more abundantly in DMSO-induced LA7 cells while the maspin protein was expressed 70-80 times more abundantly in the 106A10 cells. Northern blot analysis indicated that the mRNA expression pattern of Tm-5b and maspin parallels the protein levels detected by 2D gel electrophoresis. Antisense RNA technology was then used to confirm a biological role for the Tm-5b and maspin proteins in the process of dome formation and, presumably, mammary gland differentiation (Zucchi et al., 2001). This study demonstrates that, despite the large numbers of proteins expressed in mammalian cells, useful information can be obtained by 2D gel electrophoresis of whole cell protein lysates.

Another interesting study has examined the proteome of the mouse brain (Gauss et al., 1999). Total brain proteins were fractionated and separated using 2D electrophoresis in a large gel format. The 2D gel pattern consisted of 8,600 protein spots for the soluble protein fraction. The identity of the protein in 331 of the spots was determined by mass spectrometry and these spots were found to encode 90 different proteins (Gauss et al., 1999). This result illustrates that individual proteins can be found at multiple positions on a 2D gel. For example, the apg-2 protein made up a "spot family" that included more than 52 spots. The spot families are thought be due to protein modification, degradation or a combination of the two. Interestingly, the spot families were different when different strains or species of mice were examined (Gauss et al., 1999). In total, over 1,000 protein spots were found to vary when mice of different genetic backgrounds were examined. This finding suggests there are genetically determined differences in post-translational modifications between strains and species of mice. The presence of a protein in multiple spots can provide information about post-translation modifications but it also makes protein expression mapping very complex, particularly in mammalian systems.

Protein expression mapping in microbial systems

All of the above examples deal with complex proteomes where only a fraction of the proteome can be examined. However, protein expression mapping has also been performed with less complex bacterial systems where it is feasible to visualize the entire proteome using 2D gel electrophoresis. For example, *Haemophilis influenzae* has been the subject of several proteomics studies. This organism is appealing as a model for protein expression mapping because the entire genome has been sequenced, its proteome includes only 1742 gene products, and it is easily cultivated in defined media. The soluble protein fraction of this organism has been analyzed by 2D electrophoresis (Langen et al., 2000). A number of pre-fractionation steps were performed to enrich for low-abundance *H. influenzae* proteins that were not visible on the 2D map obtained with a crude protein extract (Langen et al., 2000). These steps included affinity chromatography on heparin, ion exchange chromatography, and hydrophobic interaction chromatography. By use of these various techniques in combination with MALDI-TOF mass spectrometry, a total of 502 different proteins were identified on the 2-D gels (Langen et al., 2000).

Gmuender et al. (2001) made use of the *H. influenzae* protein database to facilitate a functional genomics approach to examine changes in gene expression that occur upon exposure of the bacterium to the antibiotics ciprofloxacin and novobiocin. Although both of these drugs inhibit bacterial DNA gyrase, they do so by different mechanisms. Novobiocin is a coumarin antibiotic that binds to the ATP binding site of the B subunit of gyrase and inhibits the supercoiling activity of the enzyme. Because transcription is sensitive to the state of supercoiling, novobiocin may affect the transcription of many genes. In contrast, ciprofloxacin is a quinolone that binds to the A subunit of gyrase and interrupts the DNA cleavage and resealing activity of the enzyme. Failing to seal double strand breaks results in DNA damage and quinolones are known to induce DNA repair systems (Piddock et al., 1990). Thus, the two drugs may elicit different physiological responses despite binding to the same target.

A distinguishing feature of this work is that Gmuender et al. (2000) examined gene expression at the level of transcription with microarrays and at the translational level using 2D gel electrophoresis. Using parallel cultures for the transcription and translation studies permitted a direct comparison between differential RNA synthesis versus differential protein synthesis. The authors found that absolute levels of an RNA species and the corresponding protein exhibit a correlation coefficient of only 0.5. However, there was a correlation between the sign of the change of RNA and protein, i.e., if an RNA exhibits increased expression, the corresponding

protein exhibits increased expression. Therefore, the observed levels of RNA and protein are qualitatively similar but the actual magnitude of the changes is significantly different. The clear conclusion from these results is that the microarray and 2D protein gel data cannot be interpreted quantitatively. Furthermore, extensive control experiments indicated that poor reproducibility of the 2D protein gels compared to microarray hybridization is a major limiting factor for comparing transcription and translation results. An important challenge for proteomics is to achieve the reproducibility of the nucleic acid microarray experiments.

The effect of novobiocin and ciprofloxacin on *H. influenzae* gene expression was assessed at the minimum concentration of antibiotic known to inhibit bacterial growth (MIC) as well as at a ten-fold higher concentration. A common theme of both the novobiocin and ciprofloxacin data is that use of the low antibiotic concentrations for short periods of time provides the most interpretable results. Use of high concentrations for extended periods of time changes the expression of a large set of genes and these effects may be secondary to the action of the drug. Nevertheless, the changes that occur at high antibiotic concentrations are largely unique to each drug and thus provide a 'signature' for that drug. An important contribution of this work is the demonstration that two antibiotics that act on the same bacterial target but by a different mechanism results in different gene expression profiles. As the authors point out, having a set of signatures for a large number of antibiotics may be a very useful tool for understanding the mechanism of action of novel pharmaceuticals. It will be of great interest to determine if exposure of bacteria to drugs that have the same mechanism of action results in the same signature of gene expression. For example, does exposure to two different quinolones result in a similar pattern of gene expression? If these types of experiments are to be used to determine the mechanism of action of new antibiotics, the answer should be yes.

This work of Gmuender et al. (2001) clearly illustrates the value of examining large sets of proteins for differences in levels in response to drug treatment. Several toxicology studies also emphasize this point. For example, a protein expression mapping approach was used to demonstrate an association between decreased levels of a calcium-binding protein and cyclosporine A-induced nephrotoxicity when kidney samples were compared from species that were either susceptible or resistant to nephrotoxicity (Aicher et al., 1998). In addition, in a separate study, a set of peroxisomal proliferator drugs was found to result in coordinated changes in mouse liver protein expression that correlated with peroxisomal β-oxidation (Anderson et al., 1996).

The Gmuender (2001) study also illustrates the limitations in terms of the quantitative reproducibility of protein expression mapping using 2D gels. They found that 32% of the spots on the 2D gels exhibited greater than

two-fold changes in intensity when gel-to-gel reproducibility was examined using an identical sample. When a different sample that was prepared under identical conditions was examined, 39% of the spots exhibited greater than two-fold changes in intensity. Thus, protein expression mapping using 2D gels is restricted to detecting those proteins exhibiting relatively large changes in expression level. Nevertheless, the qualitative information provided by protein expression mapping is of clear utility.

3.2 Quantitative protein expression mapping

Metabolic labeling ofproteins with radioactive amino acids

The problem of quantitation in the use of 2D gel electrophoresis for protein expression mapping has been addressed by labeling of proteins either *in vivo* during cell growth or *in vitro* within the protein lysate. For example, protein expression mapping has been performed with *Saccharomyces cerevisiae* by growing cells in the presence of [^{35}S]methionine and fractionating the protein extract by 2D gel electrophoresis (Gygi et al., 1999). The identity of the proteins corresponding to each spot was determined by in-gel digestion, fractionation of the resulting peptides by liquid chromatography, and sequence analysis by tandem mass spectrometry (Gygi et al., 1999). Excising the spot from the gel and determining the amount of radioactivity using a scintillation counter, in turn, quantitated the amount of protein in a spot. The exact amount of protein in each spot was calculated from this number by comparison to protein standards of known concentration (Gygi et al., 1999).

The metabolic labeling method was used to quantitate protein expression for 128 yeast genes. Expression levels varied from 2,200 to 863,000 copies of a protein per cell (Gygi et al., 1999). This data was used to compare protein expression levels with the levels of the corresponding mRNAs. The levels of mRNA were taken from frequency tables for mRNA transcripts that were generated by serial analysis of gene expression (SAGE) (Velculescu et al., 1997). The comparison is possible because the proteome analysis was performed on the same yeast strain that was grown under the same conditions as the SAGE analysis. Correlation coefficients of 0.1 to 0.4 were calculated for protein and mRNA levels when the analysis was restricted to proteins expressed at low to moderate levels. When highly abundant proteins were added to the comparison, the correlation coefficient increased to 0.94. Thus, there is a strong correlation between mRNA and protein levels for highly expressed proteins. However, based on the average abundance levels of yeast proteins, it is estimated that the correlation coefficient for all yeast proteins or for proteins selected at random would be

less than 0.4 (Gygi et al., 1999). These results suggest that posttranslational mechanisms such as translational control or control of protein half-life play an important role in the regulation of protein expression levels in yeast (Gygi et al., 1999). These results also stress the importance of determining protein levels directly by proteomic methods in addition to measuring mRNA levels using SAGE or DNA chip technology (Gygi et al., 1999; Shalon et al., 1996; Velculescu et al., 1997).

Metabolic labeling ofproteins with stable isotopes

Another means of metabolic labeling of proteins for quantitation makes use of stable isotope labeling. For these experiments, two cell populations, such as a wild type and mutant, are studied. This approach has been used to compare protein expression levels in a *Saccharomyces cerevisiae* strain lacking the CLN2 gene and its wild-type parent (Oda et al., 1999). The CLN2 gene product is a cyclin protein that is important in regulating the G1 to S phase transition in yeast (Cross, 1995). The wild-type cells were grown in a medium containing the natural abundance of nitrogen isotopes, i.e., 99.6% ^{14}N and 0.4% ^{15}N, while the *cln2* cells were grown in the same medium enriched to >96% in ^{15}N. After cell growth, the pools were combined and a crude protein lysate was produced and fractionated by reverse-phase HPLC and PAGE (Oda et al., 1999). Individual gel spots were then excised, trypsinized and analyzed by mass spectrometry. The ratio of peptides derived from cells grown in ^{14}N versus ^{15}N could be readily determined because the mass of peptides containing ^{15}N is shifted upwards from the ^{14}N-containing peptide to produce two peaks (Oda et al., 1999) (Fig. 3.2). The ratio of the intensities of these peaks provides an accurate determination of the relative abundance of the peptides.

The ratio of peaks for peptides derived from 42 high abundance yeast proteins was examined for the wild type versus *cln2* mutant strains (Oda et al., 1999). Only two of the proteins, a peroxisomal membrane protein and S-adenosylmethionine synthase 2, exhibited significant differences in expression between the strains. The biological significance of this observation is not yet known but the study does indicate that changes of >20% in expression levels can be detected using the technique (Oda et al., 1999).

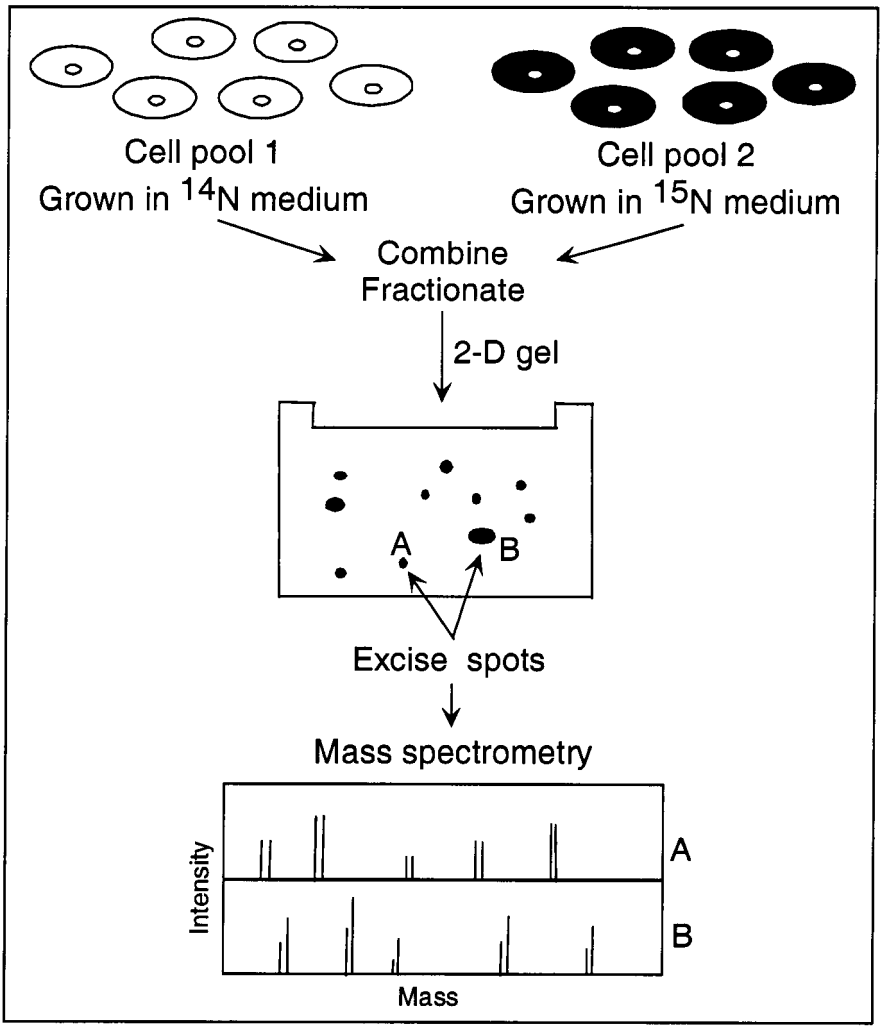


Figure 3.2. Stable isotope labeling for quantifying differential protein expression. Cell populations are grown in either ^{14}N or ^{15}N containing medium. Protein lysates are fractionated and separated by 2D gel electrophoresis. Protein spots are excised, digested with trypsin and the mass of the resulting peptides is determined by mass spectrometry. The presence of ^{15}N results in a shift and creates two peaks for each peptide. The ratio of intensities of the peaks is indicative of the relative expression levels of the proteins. Spot A contains a protein that is expressed at similar levels in both cell pools. Spot B contains a protein that is expressed at higher levels in cell pool 2. Figure adapted from Oda et al. (1999).

In vitro labeling of proteins using isotope-coded affinity tags

The use of the *in vivo* labeling methods described above is limited by the fact that the sample must be grown in the presence of the labeling isotopes. In many cases, it is not feasible to perform *in vivo* metabolic labeling. For example, for human clinical samples it is not possible to perform *in vivo* labeling and yet it is highly desirable to obtain accurate quantitative information on protein expression levels within these samples. Therefore, robust methods are needed for quantitation of protein levels in the absence of *in vivo* labeling with isotopes.

A method has recently been developed for quantitating protein expression that involves labeling proteins *in vitro* with a molecule termed an isotope-coded affinity tag (ICAT) (Gygi et al., 1999). The tag contains a biotin group for affinity purification, an isotopically coded linker, and a thiol-reactive group for attaching the tag to cysteine residues within the protein lysates (Fig. 3.3). The linker region is used to create heavy or light versions of the tag. The heavy form contains eight deuteriums in the coded linker while the light form contains no deuteriums so that there is an 8 Dalton difference in mass between the tags (Gygi et al., 1999). The experimental approach is similar to that of metabolic labeling with stable isotopes in that two cell populations are examined. However, for the affinity tag approach, the labeling is done *in vitro* with the protein lysate. The cysteines within the proteins from each mixture are first reduced and then one sample is derivatized with the heavy tag and the other sample with the light tag. The protein mixtures are then cleaved with trypsin and the resulting peptides are purified on an avidin affinity matrix. The peptides that have been derivatized with the heavy and light tags are pooled and separated by high performance liquid chromatography. The final step is analysis of the tagged peptides by mass spectrometry (Gygi et al., 1999). The peptide pairs containing heavy and light tags are chemically identical and therefore coelute from the chromatography step. In addition, because there is an 8 Dalton mass difference between the peptide pairs, they are easily resolved in the mass spectrometry analysis. The ratio of the heavy to light peptides identified by mass spectrometry is an indication of the relative ratios of the peptides, and thus the proteins, in the original samples.

The results obtained with the ICAT labeling strategy are similar, in principle, to those obtained by *in vivo* labeling with stable isotopes in that the relative ratios of proteins from different samples are obtained (Fig. 3.2). The important difference between the methods is that the ICAT-labeling procedure is performed *in vitro* on protein lysates. Therefore, the ICAT-labeling strategy can be applied to samples that cannot be labeled *in vivo.* Because the ICAT-labeling method can be used with virtually any sample, it

is likely to be used widely to quantitate protein expression differences between samples. One limitation of the method is the requirement for cysteine residues within the target proteins for labeling by the ICAT-tag. This limitation is likely to be removed as ICAT reagents with different specificities are developed.

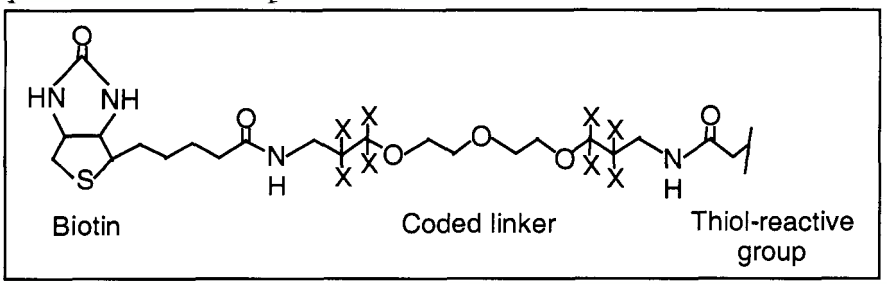


Figure 3.3. Structure of the ICAT reagent. The reagent contains a biotin affinity tag that is used to isolate ICAT-labeled peptides. The reagent also contains a linker that exists in a heavy (where X= deuterium) or light form (X= hydrogen); and a reactive group with specificity towards the thiol groups of cysteine residues. Figure adapted from Gygi et al. (1999).

Conclusions

Protein expression mapping using 2D gel electrophoresis and mass spectrometry is the experimental strategy most often associated with the term proteomics. This approach is becoming widely used because of the desire to examine protein levels directly and because the instrumentation necessary for these experiments is readily available. As discussed above, protein expression mapping is often performed in conjunction with RNA expression studies using microarray technology. However, several limitations currently impede the progress of protein expression mapping. For example, the sensitivity and reproducibility of 2D gel electrophoresis does not match that of microarray technology (Gmuender et al., 2001; Gygi et al., 2000). In addition, quantitative analysis is difficult because the same protein can be found in many different spots on a 2D gel and because mass spectrometry cannot be used for quantitation without a reference point provided by isotope labeling (Gauss et al., 1999; Gygi et al., 1999). However, recent advances such as the *in vitro* ICAT-labeling strategy may solve many of these problems. The ability to fractionate, identify and quantitate proteins from complex samples by coupling ICAT-labeling with mass spectrometry has the potential to improve both throughput and reproducibility because fewer manipulations are required. In addition, the ICAT method could lead to improved detection of proteins that are present in low amounts in samples because, in contrast to 2D gel electrophoresis, any amount of starting material can be used.

This page intentionally left blank.

Chapter 4

HIGH-THROUGHPUT CLONING OF OPEN READING FRAMES

Both structural and functional genomic studies are critically dependent on efficient cloning of the open reading frames (ORFs) identified by genome sequences. The ORFs must be cloned into plasmid vectors that permit large-scale expression or facilitate functional analysis of the encoded proteins. Because all proteins are not expressed equally well from a single system, it is necessary to obtain constructs whereby the gene of interest is present in multiple protein expression systems. For example, it would be useful to test expression of a gene from the T7, *tac, trc* or λ P_L promoters to determine which system is optimal for expression. In addition, for functional analysis, it would be desirable to place the ORFs into vectors used for phage display, two-hybrid analysis or green fluorescent protein fusions. Using conventional cloning techniques, the necessary construction time and cost of such vector sets is prohibitive. However, new methods have been developed to facilitate the direct cloning of PCR products.

4.1 Topoisomerase-based cloning

Vaccinia virus topoisomerase I adapted vectors

As discussed above, alternative recombinant DNA techniques are necessary to efficiently generate genome-scale clone sets. One alternative exploits the ability of the Vaccinia virus DNA topoisomerase I to both cleave and rejoin DNA strands with high sequence specificity (Shuman, 1992a; Shuman, 1992b). In the reaction, the enzyme recognizes the sequence 5'-CCCTT and cleaves at the final T whereby a covalent adduct is formed between the 3' phosphate of the cleaved strand and a tyrosine residue in the enzyme (Fig. 4.1). The covalent complex can combine with a heterologous acceptor DNA that has a 5' hydroxyl tail complementary to the sequence on the covalent adduct to create a recombinant molecule (Shuman, 1994).

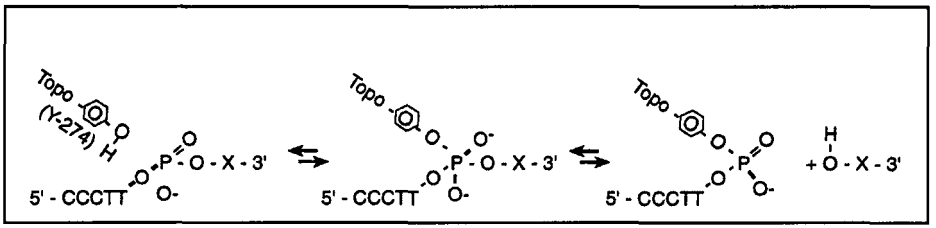


Figure 4.1. Vaccinia virus topoisomerase I reaction with DNA to form a covalent adduct. Figure adapted from Heyman et al. (1999).

Topoisomerase I-based cloning exploits the reaction described above to join DNA fragments containing 5' hydroxyl groups to acceptor plasmids to which the Vaccinia topoisomersase I enzyme is covalently attached. Because the joining reaction only occurs if the incoming DNA has a free 5' hydroxyl, it is ideal for cloning PCR products because these do not possess 5' phosphates as the oligonucleotide primers used for amplification do not have *5'* phosphate groups (Shuman, 1994). The efficiency of this cloning method has recently been demonstrated by its use in the cloning of 6035 ORFs from *Saccharomyces cerevisiae* into a plasmid vector for protein expression in yeast as well as a vector for expression in mammalian cells (Heyman et al., 1999). One drawback of the original topoisomerase cloning method is that the PCR product can insert in either orientation. The cloning system has recently been improved to allow directional cloning of PCR products (Fig. 4.2). The only requirement for PCR primer design for this system is to include the sequence 5'-CACC at the 5' end of the PCR product. The 5'-CACC sequence is complementary to a sequence on the 5' side of the plasmid-topoisomerase I adduct and as such controls the orientation of insertion of the PCR product (Fig. 4.2).

The topoisomerase cloning method greatly increases the efficiency of cloning and thus permits the systematic insertion of large numbers of ORFs into a plasmid vector. For structural and functional genomics studies, however, it is necessary to insert ORFs into many different types of vectors including those that facilitate high levels of protein expression or those that enable the fusion of tag sequences to the ORF. One approach would be to use topoisomerase-based cloning to insert the PCR product containing the gene of interest into a number of different plasmids. However, the high error rate of thermostable polymerases requires that the sequence of each PCR-generated fragment be verified. Thus, it would be necessary to sequence the same gene in each cloning vector. Clearly, this would be an expensive and time-consuming process if repeated for all of the ORFs in a genome.

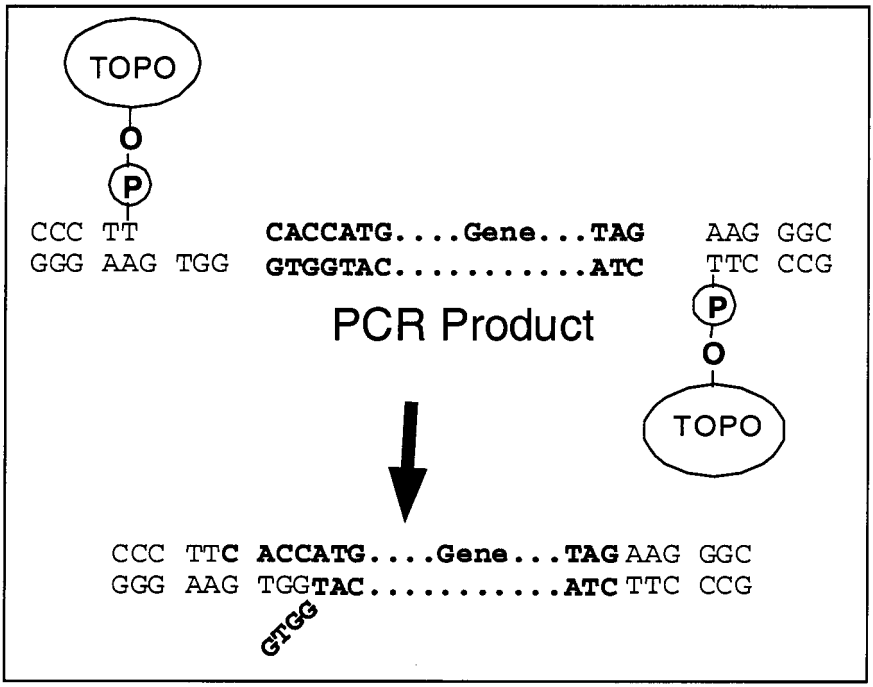


Figure 4.2. Schematic illustration of directional cloning of PCR products. The sequence 5'CACC is required at the 5' end of the PCR product for directional topoisomerase-mediated cloning. In the example shown, the 5'-CACC sequence is appended immediately 5' of the ATG start codon of the gene to be inserted.

4.2 Univector plasmid-fusion system

Cre-lox mediated recombination

An alternative to repeated cloning of PCR products is a recombination-based approach developed by Liu et al. (1998) to permit the cloning of a PCR product into a plasmid and the rapid conversion of the plasmid to a number of different expression systems without the necessity of cloning the PCR product multiple, independent times. The method, termed the univector plasmid-fusion system (UPS), involves the insertion of the PCR product into a particular type of plasmid, called the univector, which can then be placed under the control of a variety of promoters or fused in-frame to various tag sequences. The system is based upon plasmid fusion using the Cre-*lox* site-specific recombination system of bacteriophage P1 (Sternberg et al., 1981). The Cre enzyme is a site-specific recombinase that catalyzes recombination between two 34 base pair (bp) *loxP* sequences and is involved in the resolution of dimers formed during replication of the

circular P1 chromosome (Sternberg et al., 1981). Cre can perform the recombination reaction both *in vivo* and *in vitro* (Abremski et al., 1983; Liu et al., 1998). The univector cloning system is illustrated in Figure 4.3. The pUNI plasmid is used for the initial cloning of PCR products. The pHOST plasmid contains the appropriate promoter sequences or tag sequences for creating fusion proteins. The recombinant protein expression construct is made by fusion of the pUNI and pHOST plasmids mediated by Cre-*loxP* site-specific recombination (Liu et al., 1998).

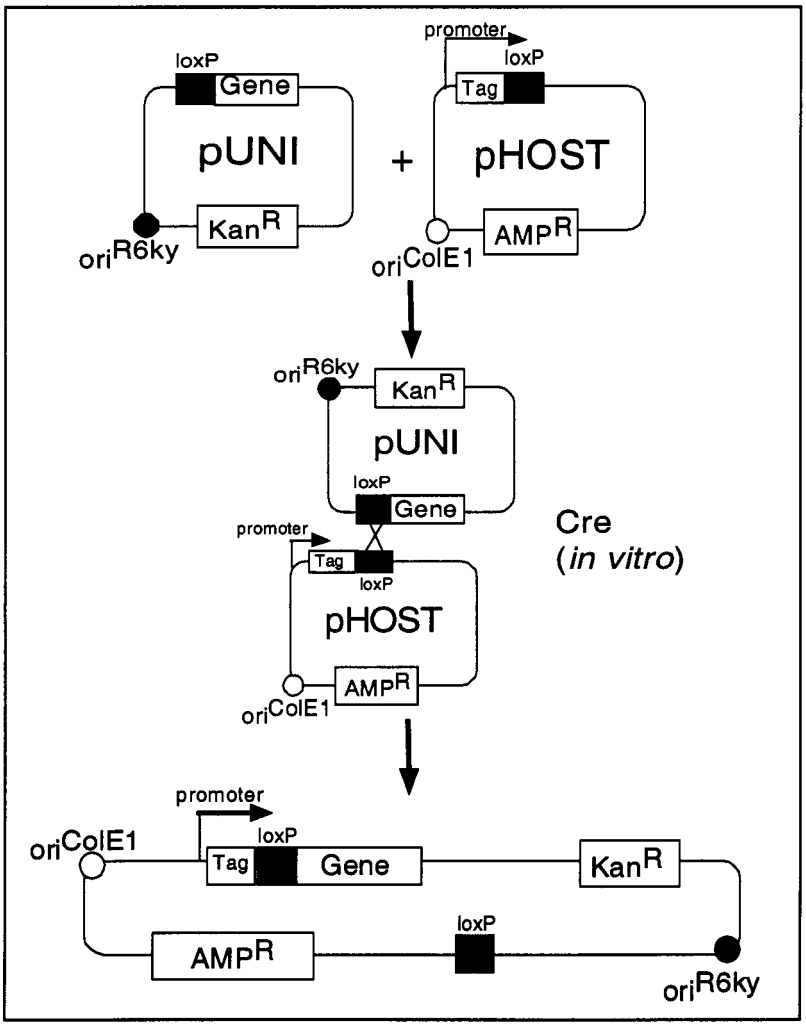


Figure 4.3. Univector plasmid fusion system. Cre-*loxP* mediated site-specific recombination fuses the pUNI and pHOST plasmids at the *loxP* site. As a result, the gene of interest is placed under the control of the pHOST promoter and fused to any Tag sequences present in the pHOST plasmid.

One of the advantages of the univector system is that a number of different pHOST vectors can be fused to the pUNI plasmid containing the gene of interest. For example, the pHOST vector could allow the generation of protein fusions to glutathione-S-transferase (GST), a poly-histidine sequence, green fluorescent protein or any tag protein of interest. Importantly, the Cre- *loxP*- mediated fusion does not necessitate that the gene of interest be sequenced again because PCR has not been employed to make the fusion.

Additional features have been included in pUNI and pHOST plasmids to facilitate the construction of fusion plasmids (Liu et al., 1998). For example, the pUNI vector contains a conditional origin of replication derived from the R6Kγ plasmid that is dependent on the action of the pir gene product (Metcalf et al., 1994). The pir gene encodes the essential π replication protein of R6Kγ and therefore replication of the pUNI plasmid will only occur *E. coli* host strains that contain the pir gene (strains normally used for recombinant DNA manipulations do not contain the pir gene) (Metcalf et al., 1994). Thus, the original PCR product is inserted in the pUNI vector and maintained in a pir^+ strain of *E. coli.* The pUNI plasmid also encodes the kanamycin resistance gene from the Tn5 transposable element. In contrast, the pHOST vector contains the ColE1 origin of replication which functions in all commonly used strains of *E. coli* and also contains the *bla* gene that encodes ampicillin resistance. After the Cre-mediated fusion reaction, the plasmid is introduced into a strain of *E. coli* lacking the pir gene and the transformed bacteria are spread on agar plates containing both kanamycin and ampicillin. Under these selective conditions, only bacteria containing fusion plasmids will be propagated. This feature greatly facilitates the rapid generation of pUNI-pHOST fusion plasmids (Liu et al., 1998).

As seen in Fig. 4.4, topoisomerase-based cloning and the univector plasmid fusion system are compatible approaches for cloning large sets of open reading frames. Topoisomerase-based cloning is an efficient method to create the initial clone set in the pUNI vector using PCR products generated for each of the open reading frames. Once the clone set has been established in the univector, a number of different functional vectors can be created for each of the ORFs by Cre- *lox* recombination with various pHOST vectors (Liu et al., 1998).

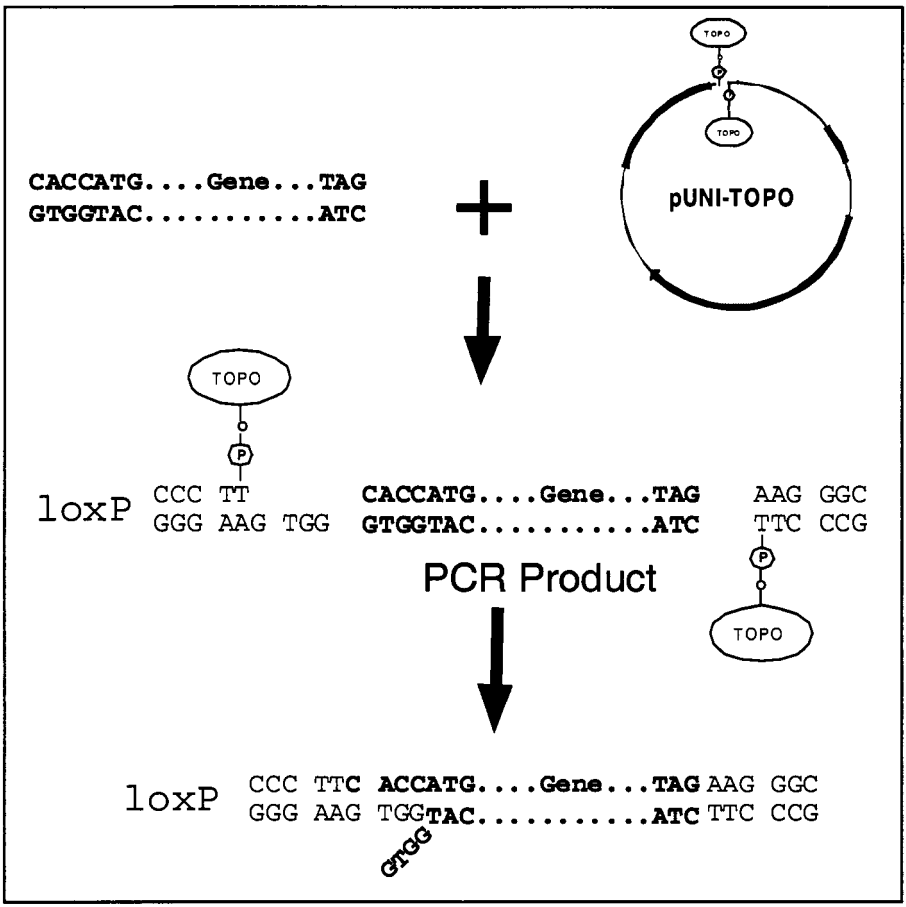


Figure 4.4. Schematic illustration of directional topoisomerase cloning of PCR products into the pUNI vector. The PCR product to be cloned has the sequence 5'-CACC appended at the 5' end to direct the orientation of cloning. The Vaccinia virus topoisomerase I enzyme forms a covalent adduct with the cloning vector to create a cloning competent plasmid construct. The *loxP* site is 5' to the insertion site. The vector and PCR product are designed to fuse the ORF in-frame with *loxP*.

4.3 Bacteriophage λ att recombination-based cloning

Cloning and gene transfer

The topoisomerase and Cre-*lox*-based cloning systems described above may not be ideal for all experimental situations. However, alternative

recombination-based cloning systems are available. One such system is based on the bacteriophage λ site-specific recombination system. Some phages, such as the λ phage, undergo both a lytic and a lysogenic cycle. In the lysogenic cycle, the phages do not multiply but, instead, their DNA integrates into the host chromosome (Ptashne, 1992). Lysogeny occurs when, immediately after infection, the λ DNA circularizes and the Int protein promotes the integration of the circular DNA into the chromosome (Fig. 4.5). The Int protein is a recombinase that catalyzes the site-specific recombination between an attachment sequence on the phage DNA (called attP for attachment phage) and an attachment sequence on the *E. coli* chromosome (called attB for attachment bacteria). An *E. coli* host protein, Integration Host Factor (IHF), is also essential for the integration of DNA (Landy, 1989). The integration reaction is highly sequence specific and is conservative, i.e., there is no net gain or loss of nucleotides. The site-specific recombination reaction is a type of non-homologous recombination because the *attB* and *attP* sites are mostly dissimilar. The sites have a common core sequence of 15 base pairs-GCTTTTTTATACTAA. Because the region of homology is small, the reaction would not occur without the Int protein, which recognizes both the *attB* and *attP* sites. The sites resulting from recombination between *attB* and *attP* are called *attL* and *attR* (Fig. 4.5)(Landy, 1989).

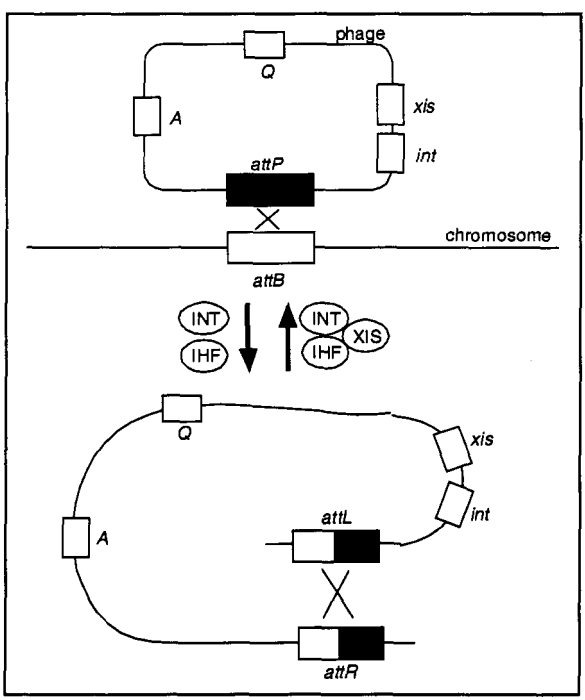


Figure 4.5. Integrative and excisive λ phage recombination pathways. Integration is catalyzed by the λ Int protein in a reaction that also requires the *E. coli* IHF protein. Recombination occurs within a common core sequence of 15 base pairs. The excision reaction requires the λ Xis protein in addition to Int and IHF.

Phage λ DNA will remain in the prophage state until the host cell DNA is damaged (Ptashne, 1992). DNA damage triggers transcription of the λ *int* and *xis* genes, among others. The Int and Xis gene products result in excision of the λ DNA from the bacterial chromosome. Excision results from recombination between the *attL* and *attR* sites to regenerate the *attB* and *attP* sites on the bacterial and phage chromosomes. The excision reaction is not simply the reverse of the integration reaction in that the Xis protein is required in addition to the Int and IHF proteins (Landy, 1989). Thus, the direction of the recombination reaction can be controlled with the appropriate combination of proteins (Fig. 4.5).

The λ recombination cloning system is a method whereby a DNA fragment flanked by *att* recombination sites can be combined *in vitro* with a vector that also contains recombination sites and incubated with integration proteins to transfer the DNA fragment into the vector (Hartley et al., 2000). This system has been termed recombinational cloning (RC) (Hartley et al., 2000). The system can be used to directly clone PCR fragments into a vector in the absence of DNA ligase, analogous to topoisomerase-based cloning. The reaction used to insert PCR fragments is *attB* + *attP* > *attL* + *attR,* which is similar to insertion of the λ prophage and is catalyzed by the addition of the Int and IHF proteins. The substrates for the reaction are a PCR fragment containing an *attB* site at each end and a plasmid containing a selectable marker flanked by *attP* sites (Fig. 4.6). In contrast to λ integration, this reaction utilizes two *attB* sites and two *attP* sites. Furthermore, the *att* sites are mutated such that *attB1* will recombine with *attP1* but not with *attP2.* The engineered differences in *att* sites permit directional cloning of PCR products into the vector (Fig. 4.6) (Hartley et al., 2000). Obtaining a cloned PCR fragment is facilitated by the presence of the F-plasmid encoded *ccdB* gene, which inhibits the growth of *E. coli,* between the *attP* sites on the recipient plasmid (Bernard and Couturier, 1992). The plasmid is maintained in an *E. coli* strain that is resistant to the effect of *ccdB.* The products of the recombination reaction are introduced into a normal strain of *E. coli* whose growth is inhibited by the CcdB protein. Therefore, only those *E. coli* cells containing plasmids that have undergone recombination with the PCR fragment to eliminate the *ccdB* gene will grow (Hartley et al., 2000).

The end result of cloning PCR fragments using RC is a vector containing a gene flanked by *attL* sites (Fig. 4.6). This plasmid is termed an Entry Clone because it can be used to generate a wide variety of functional vectors by an additional recombination reaction (Hartley et al., 2000). The reaction used for vector conversion is *attL* + *attR* > *attB* + *attP,* which is similar to excision of the λ prophage by the Int, Xis and IHF proteins (Landy, 1989). The gene within the Entry Clone is transferred to a destination vector that contains the desired transcriptional promoters and protein tags by incubating the two vectors with the Int, Xis and IHF proteins.

The reaction differs from excision of the λ chromosome because the Entry Clone contains two *attL* sites and the destination vector contains two *attR* sites (Hartley et al., 2000). The *att* sites are mutated to ensure recombination only occurs between *attL1* and *attR1* and between *attL2* and *attR2*. The recombination reaction proceeds through a cointegrate molecule that is resolved to create a destination vector containing the gene of interest with the desired promoter and tag sequences (Fig. 4.6).

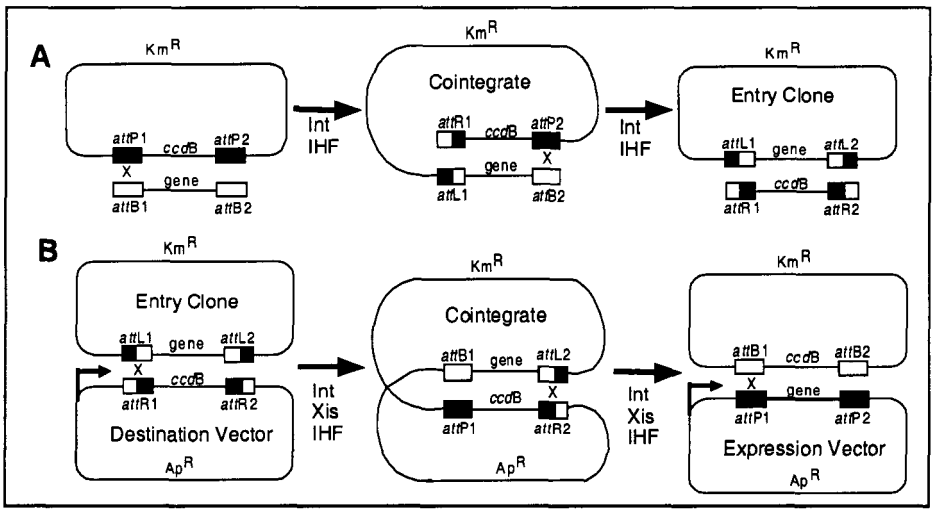


Figure 4.6. Recombinational cloning (RC). **A.** Cloning of PCR products using the *attB* + *attP* > *attL* + *attR* reaction catalyzed by Int and IHF. The result is an Entry Clone that can be used to create functional vectors. **B.** Conversion of an Entry Clone to a functional vector using the *attL* + *attR* > *attB* + *attP* reaction catalyzed by Int, Xis and IHF. A wide variety of functional vectors can be constructed by using a Destination Vector with the appropriate promoters and tags.

The λ recombination cloning system is a versatile method for high throughput cloning of genes. The use of the *attB* + *attP* > *attL* + *attR* reaction to clone PCR products and create Entry Clones is analogous to the topoisomerase-based cloning method. In addition, the use of the *attL* + *attR* > *attB* + *attP* reaction to transfer genes from the Entry Clone to a destination vector is analogous to the univector plasmid fusion system (Hartley et al., 2000; Liu et al., 1998). A difference between the systems is that the univector fuses with a host vector while in the RC system the fusion of plasmids creates a cointegrate that is resolved to precisely transfer the gene of interest to the destination vector (Hartley et al., 2000; Liu et al., 1998). However, a version of the univector cloning system has been developed that utilizes both a *loxP* site for plasmid fusion by Cre and an RS site for cointegrate resolution by the R recombinase of yeast (Araki et al., 1992; Liu

et al., 1998). Thus, precise transfer of an open reading from the starting vector to the final, functional vector is possible with both systems.

The RC system has been used recently for some large scale cloning projects. For example, the RC system is being used to clone thousands of genes from the worm *Caenorhabditis elegans* to create vectors for use in protein expression and analysis of protein-protein interactions using the yeast two-hybrid system (Walhout et al., 2000). In addition, greater than 100 human cDNAs have been cloned to create a set of Entry Clones using the RC system. These clones were then converted to green fluorescent protein fusions using an appropriate destination vector and the cellular localization of the proteins encoded by the cDNAs was determined (Simpson et al., 2000).

4.4 *In vivo* recombination-based cloning in yeast

Cloning of PCR products by transformation

If neither of the above methods is suitable for a particular cloning application, yet another method is available. The λ Int recombination system described above utilizes *in vitro* recombination followed by transformation of the products into *E. coli* cells. An analogous method that utilizes the highly efficient homologous recombination machinery of yeast has been used to clone greater than 99% of the *Saccharomyces cerevisiae* genes (~6,000) (Uetz et al., 2000). This was accomplished using a two-step PCR procedure. A set of ~6000 primer pairs were used to amplify the ~6000 ORFs from the *S. cerevisiae* genome. Each forward primer contained a sequence unique to the ORF as well as a 22 base pair (bp) sequence at the 5' end that was common to all of the forward primers. Similarly, the reverse primer contained a sequence unique to each ORF as well as a 20 bp sequence at the 5' end that was common to all of the reverse primers. Each of the 6000 ORFs was then amplified to generate the initial set of PCR products (Fig. 4.7).

The second set of PCR reactions was performed using primers complementary to the 22-bp and 20 bp sequences appended on the initial forward and reverse primers. In addition to the complementary sequences, these forward and reverse primers contained an additional 50 base pairs of sequence homologous to sequences flanking a cloning site in the yeast vector to be used as the recipient plasmid for cloning (Hudson et al., 1997; Uetz et al., 2000). The resulting PCR products therefore contained 70 base pairs of sequence at each end that was homologous to 70 base pairs on either side of the cloning site in the recipient vector (Fig. 4.7).

Each of the ~6000 PCR products was then co-transformed into yeast along with the recipient vector that had been linearized using a restriction enzyme that digests the plasmid at the desired cloning site. The 70 bp of homologous flanking sequence on each end of the PCR products is sufficient for the yeast homologous recombination system to act upon and insert the PCR product into the vector (Hudson et al., 1997; Ma et al., 1987).

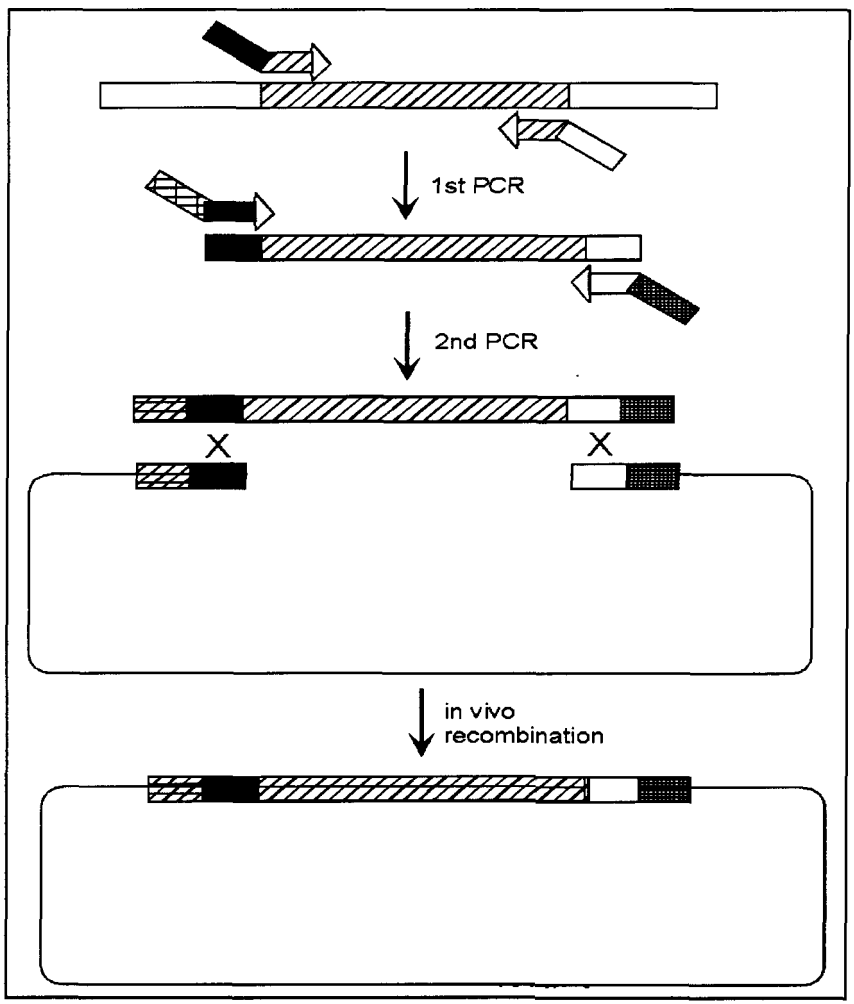


Figure 4.7. In vivo recombinational cloning in yeast. Two successive PCR reactions are performed. Each set of primers contains 5' flanking sequences that are eventually used for homologous recombination with vector sequences after transformation of yeast with the 2nd PCR fragment and the linearized vector.

4.5 Advantages and disadvantages of recombinational cloning systems

All of the recombinational cloning methods are efficient methods for cloning PCR products into a desired set of vectors. The main advantage of topoisomerase-based cloning, besides ease of use, is the minimal requirement for extra sequences being appended to the 5' end of the primers used for the amplification of individual ORFs. Efficient directional cloning of PCR products can be achieved with only 4 bp appended to the 5' end of the forward primer and no additional nucleotides added to the reverse primer. In contrast, cloning of PCR products using the λ recombination reaction requires an additional 25 bp appended to both the forward and reverse primers (Hartley et al., 2000).

Cloning of PCR products using yeast recombination has an advantage in the relative ease of the procedure. The necessity of having >50 bp of sequence at each end of the PCR product that is homologous to sequence in the vector, however, is cumbersome. The homology requirement necessitates the two sets of PCR reactions. Performing multiple PCR reactions increases the probability that the cloned gene will contain a mutation. In addition, the primers used for the initial PCR reaction contain 20-22 bp of sequence in addition to the sequence complementary to the individual ORF, which greatly increases the cost of primer synthesis.

The most efficient strategy may be to use the topoisomerase cloning method in conjunction with the Cre-*lox* or λ recombination systems. In this way, PCR products can be efficiently inserted into the Univector or Entry Clone with minimal additional sequences added to PCR primers. Once the genes are inserted into the starting vectors they can rapidly be converted to functional vectors using either Cre-*lox* or λ recombination.

Chapter 5
PROTEIN-PROTEIN INTERACTION MAPPING: EXPERIMENTAL

Protein-protein interactions play key roles in the functioning of cells. For example, signal transduction is critically dependent on a cascade of protein-protein interactions that occur in response to outside stimuli. In addition, molecular machines composed of many different proteins that interact in a coordinated fashion carry out essential cellular functions such as DNA replication and mRNA transcription. Therefore, defining the physical interactions between proteins is an important step towards understanding the function of each gene in a genome. Defining protein interactions on a genome wide scale can establish networks of interacting proteins, which can provide important clues as to the function of a gene product. For instance, if a protein of unknown function were found to interact with a cluster of proteins whose function is known, it would suggest the protein is involved in the same function (Schwikowski et al., 2000).

Identifying protein-protein interactions can also provide new insights into the mechanism of a biological process by providing detailed knowledge of the binding partners within a complex. For example, even if a group of proteins are known to be involved in a biological process, one does not know how or if these proteins interact in complexes to perform the function. Protein-protein interaction mapping data can detail the protein interactions and thereby provide precise information on the organization of complexes.

Finally, the networks of protein-protein interactions defined by these studies provide a global view of how cellular processes are coordinated. As described below, protein interaction networks from *S. cerevisiae* indicate that interactions occur not only between proteins involved in certain cellular processes but also between proteins involved in different processes. These links may be crucial for sharing information and thereby coordinating global responses to environmental stimuli.

5.1 Yeast two-hybrid system

Overview of methodology

The most popular method for establishing genome-wide protein-protein interaction maps has been the yeast two-hybrid system; a genetic assay based on the modular properties of site-specific transcriptional activators (Fields and Song, 1989). Hybrid proteins composed of a DNA-binding domain fused with a protein X and a transcriptional activation domain fused with a protein Y is produced in yeast. If protein X and protein Y interact, it reconstitutes the transcription factor and leads to the expression of a reporter gene (Fig. 5.1). The reporter gene is chosen based on the ease of assaying its product by the growth of yeast on defined media. The DNA-binding domain fusion is frequently referred to as the "bait" while the activation domain fusion is the "prey".

The first two-hybrid assay measured the interaction between two yeast proteins involved in regulating the SUC2 gene, Snf1 and Snf4, by expressing them as fusion proteins (Fields and Song, 1989). One construct contained the DNA-binding domain of the Gal4 transcription factor fused to the amino terminus of Snf1 and the other contained an activation domain from Gal4 fused to the amino terminus of Snf4. A binding site for the Gal4 DNA binding domain was placed upstream of the *lacZ* reporter gene from *E. coli,* which encodes β-galactosidase. The interaction of Snf1 and Snf4 recruited the activation domain to the DNA-binding domain and activated transcription of the *lacZ* gene (Fields and Song, 1989). Gene activation could be monitored because fl-galactosidase production causes a colony to turn blue on X-Gal indicator plates.

Several versions of the yeast two-hybrid system have been developed in an attempt to improve the reporter assay and the ease of use of the system. Many of these systems make use of the fusions to the Gal4 DNA-binding and transcriptional activation domains as described above. An alternate version utilizes a fusion of the *E. coli* LexA DNA-binding domain to the amino terminus of the protein of interest to create the bait. This system also requires a LexA binding site placed upstream of the reporter gene. A prey construct consisting of a Gal4 activation domain or another acidic activation domain such as B42 derived from *E. coli* sequences can be used with either the Gal4 or LexA DNA-binding fusion bait constructs (Brent and Finley, 1997).

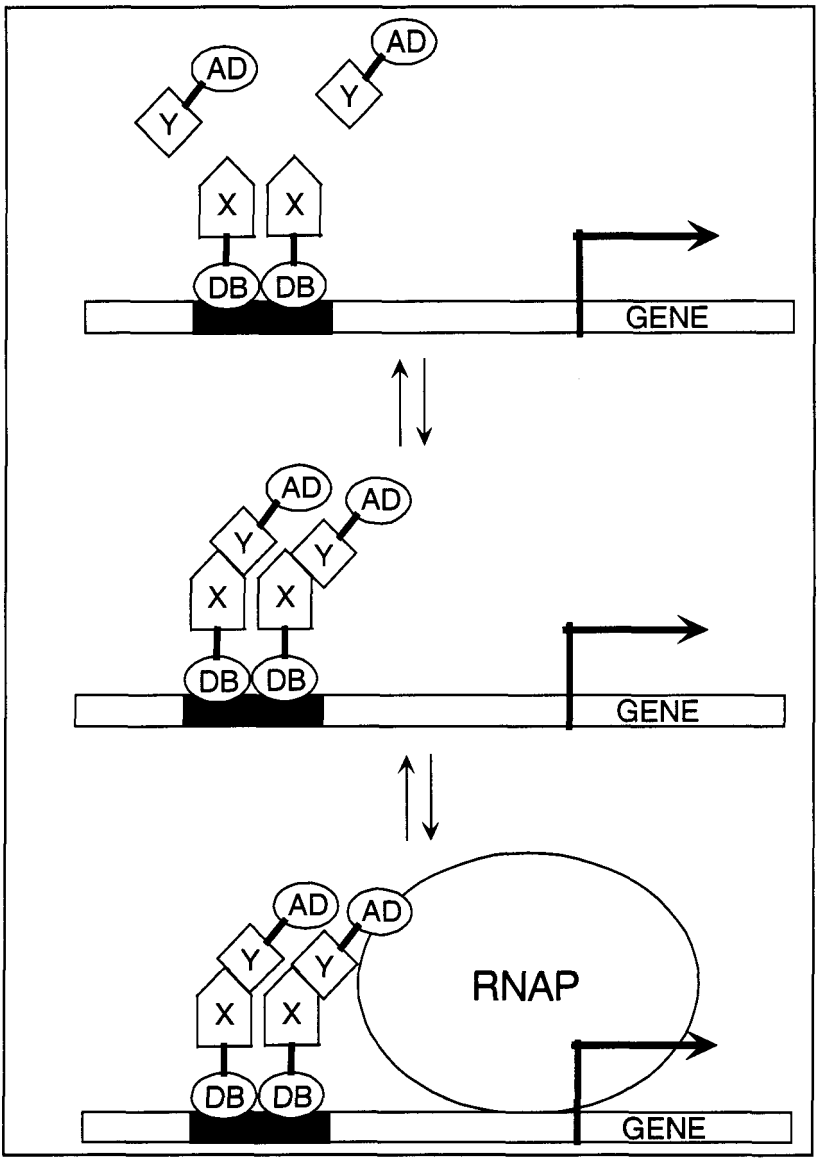


Figure 5.1. Yeast two-hybrid system. Interaction of proteins X and Y upstream of a reporter gene leads to transcriptional activation. Protein X is part of a fusion protein that binds to a site on DNA upstream of the reporter gene by means of a DNA binding domain. Protein Y is part of a fusion protein that contains a transcriptional activation domain. Interaction of proteins X and Y places the activation domain in the vicinity of the reporter gene and stimulates its transcription.

A major challenge in the use of two-hybrid systems is the elimination of false positives. These clones result from activation of reporter

gene transcription independent of specific binding between the bait and prey fusion proteins. For example, any bait protein that self-activates transcription will be a false positive. However, self-activating baits can be easy isolated and eliminated by screening the bait construct alone for reporter gene activation. A high background of false positives is also known to occur for certain reporter systems (James et al., 1996). To avoid this problem, recent versions of the two-hybrid system use multiple reporter genes with selectable or screenable phenotypes to monitor gene activation. In one version, the reporter genes include HIS3, ADE2 and *lacZ* (James et al., 1996). In addition, this system utilizes different Gal4-responsive promoters, with the HIS3 gene under the control of the GAL1 promoter, the ADE2 gene under GAL2 control and the *lacZ* gene under the control of the GAL7 promoter. Protein-protein interactions are identified by growth of yeast on agar plates lacking histidine. False positives are then eliminated using colony color screens to score for the adenine and *lacZ* markers (James et al., 1996). The use of multiple reporter genes under different promoter control is reported to provide high levels of sensitivity with low background levels of false positives (Brent and Finley, 1997; James et al., 1996).

The most established use of the two-hybrid method has been to isolate new proteins from activator domain libraries that interact with LexA or Gal4 fusion baits. The activator domain libraries can consist of cDNA or fragmented genomic DNA inserted at the C-terminus of the activator domain. New interacting proteins are identified by introducing the activator domain library into the yeast strain containing the protein of interest fused to the DNA binding domain and isolating colonies using the reporter systems described above (Bai and Elledge, 1997). This approach has been used by many laboratories and has resulted in the identification of many new interactions (Schwikowski et al., 2000).

5.2 Analysis of genome-wide protein-protein interactions in yeast

High-throughput yeast two-hybrid screens

The availability of the complete sequence of *Saccharomyces cerevisiae* has motivated attempts to map protein-protein interactions on a genomewide basis by the two-hybrid method, using gene sets consisting of all of the yeast open reading frames (ORFs) amplified and cloned individually (Hudson et al., 1997; Uetz et al., 2000). Although large-scale two hybrid experiments have been performed on other organisms, yeast has been the target of multiple studies and will be discussed first. Two types of two-hybrid experiments have been performed on a large scale. In the array method, yeast clones containing individual ORFs cloned as fusions to the

Gal4 DNA binding domain or activation domain are arrayed onto a grid and the reciprocal fusions are screened individually against the array to identify interacting clones. In the comprehensive library screening method, a set of cloned ORFs are pooled to create a library of fusions and then individual ORF fusions are mated against the library to identify interacting clones.

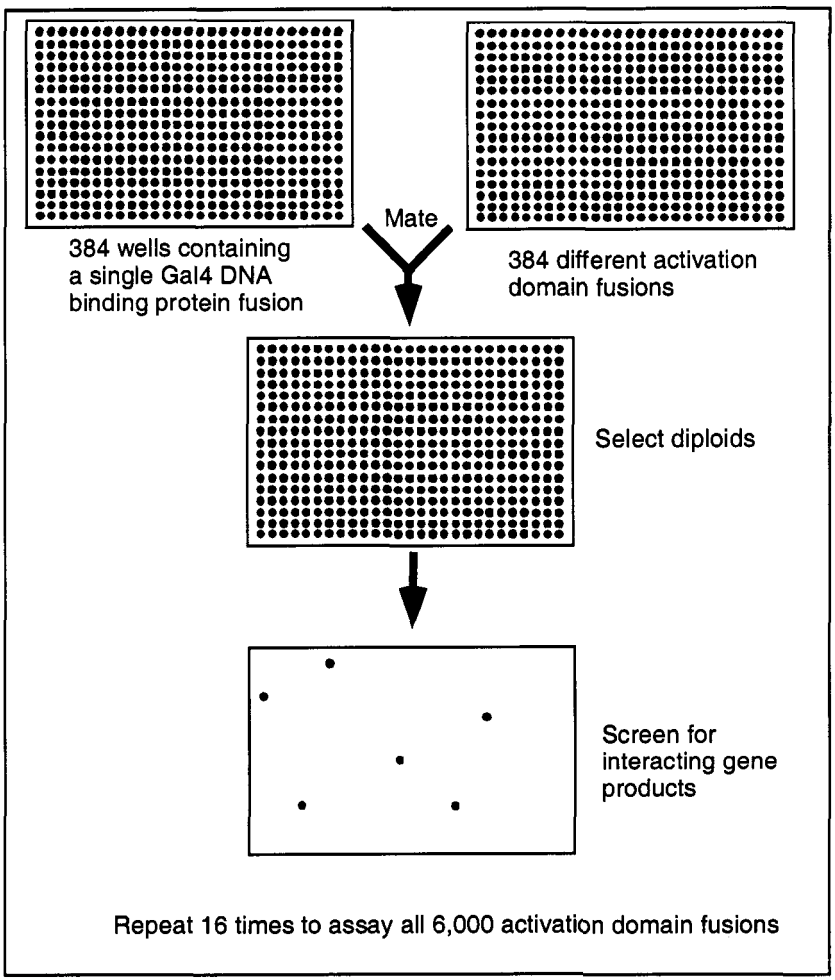


Figure 5.2. High-throughput mating assay for two-hybrid protein interaction screening. Yeast strains containing individual bait and prey clones are combined in a well and allowed to mate. Diploids are then selected and scored for a protein-protein interaction using the selection provided by the transcriptional reporter gene.

In the first large-scale array experiment, a set of approximately 6,000 genes were individually cloned as fusions to the Gal4 activation

domain (Hudson et al., 1997) and transformed into a two-hybrid reporter strain. Each strain was then inoculated into a well of a 384-well microassay plate. Sixteen such plates contained the ~6,000 strains and constituted a living protein array (Uetz et al., 2000). A set of 192 yeast genes were then individually fused to the Gal4 DNA-binding domain and transformed into a reporter strain of the opposite mating type as the activation domain clone set. Each of the 192 bait strains were then mated to each of the ~6,000 prey strains on the array to systematically screen for protein-protein interactions (Uetz et al., 2000) (Fig. 5.2). It was found that 87 of the 192 DNA-binding domain fusions participated in a protein-protein interaction, providing a total of 281 interacting pairs.

Each of the 6,000 individually cloned Gal4 activation domain fusions were collected into a single pool for the initial comprehensive library screen (Uetz et al., 2000). The set of ~6,000 yeast genes were then individually fused to the Gal4 DNA-binding domain. The pooled activation domain fusions were mated individually to each of 6,000 unique Gal4 DNA binding protein fusions. Potential interactors were identified using reporter genes and 12 clones from each mating that yielded interactors were sequenced to identify the activation domain fusion. A total of 817 ORFs were found to participate in 692 protein pairwise interactions (Uetz et al., 2000). It is of interest to note that 45% of the 192 proteins assayed were found to interact in the protein array experiment while only 8% of the 5,345 potential ORFs in the high-throughput screen using the 6,000-pooled prey constructs were found to interact. Thus, the array screen, although of low throughput, generates a proportionally higher number of interactors. It is possible that pooling all of the activation domain clones for the high-throughput experiment may select against interactions that involve cells with reduced growth rates or mating ability (Uetz et al., 2000). Regardless, the result does indicate that the set of interactors identified by the two-hybrid screen is critically dependent on how the method is implemented.

A high-throughput, comprehensive library screen has also been performed by Ito et al. (2001) in which a DNA-binding domain fusion and an activator domain fusion was constructed for each of the ~6,000 yeast ORFs (Ito et al., 2001). The DNA-binding domain and activator domain fusions were transformed into yeast reporter strains of opposite mating type and pooled in sets of 96 ORF fusions per pool to create 62 pools each for the DNA-binding fusions and activation domain fusions. To examine all possible binding interactions, 3,844 (62 x 62) mating reactions were performed between the DNA binding domain and activation domain pools. After mating, diploids containing interacting proteins were selected using multiple reporter genes and inserts from both the DNA binding and activation domain fusions were amplified by PCR and sequenced to determine the identity of the interacting clones. A total of 3,268 yeast proteins were found to participate in 4,549 pairwise interactions (Ito et al.,

2001). When the authors considered only those interactions that were identified at least three times, a core group of 806 interactions among 797 proteins emerged (Ito et al., 2001).

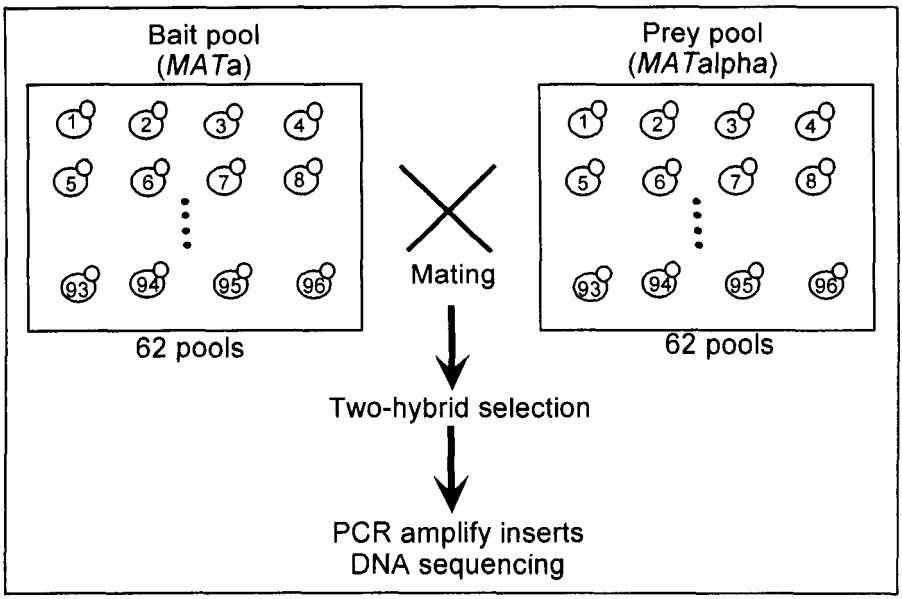


Figure 5.3. Systematic mating of yeast two-hybrid bait and prey pools. Each yeast ORF was cloned individually into both as a DNA binding domain fusion (bait) and activation domain fusion (prey). The bait fusions were introduced into a MATα strain and the prey fusions were introduced into a MATα strain. The bait and prey fusions were pooled in sets of 96 clones to generate a total of 62 pools of each. The pools were systematically mated (62 x 62) in a total of 3844 crosses. Interacting clones were selected and the bait and prey inserts were PCR amplified and sequenced to determine their identify. Figure adapted from Ito et al. (2001).

A comparison of the interactions identified from the high-throughput library screens described above reveals a surprisingly small amount of overlap, with only about 20% of the interactions in common (Ito et al., 2001; Uetz et al., 2000). The lack of overlap suggests that the library screen experiments are not saturating. This is not surprising in that the use of pools combined with DNA sequencing of selected clones is unlikely to sample a significant fraction of the potential interactions. A screen of every possible ORF by ORF interaction would require testing 6000 x 6000 = 3.6 x 10^7 pairwise combinations. Given that many proteins are known to have multiple interaction partners, the number of actual interactions could be much larger.

Computationally-directed two-hybrid screen

Further indication that the *S. cerevisiae* genomewide, two-hybrid library screens provide an underestimate of total interactions is provided by a computationally directed screen focused on identifying only those interactions mediated by a coiled coil protein motif (Newman, 2000). Coiled coils are a protein interaction motif consisting of two or more α helices that wrap around each other (Cohen and Parry, 1994). Sequences capable of forming coiled coils are characterized by a simple repeat pattern that has lead to the development of accurate computer programs that identify coiled-coil motifs from the primary amino acid sequence of proteins (Berger et al., 1995; Wolf et al., 1997). Use of such a program to identify coiled coils within yeast ORFs predicted approximately 300 proteins with two-stranded coiled coils and 250 proteins with three-stranded coiled coils encoded in the genome (Newman, 2000). Thus, approximately 1 in 11 yeast proteins are predicted to contain a coiled coil. Newman et al. (2000) examined interactions among 162 of the putative coiled coil regions using the two-hybrid system. A total of 162 x 162 = 26,244 pairwise tests identified 213 interactions involving 100 coiled coil motifs derived from 77 different proteins.

Strikingly, none of the interactions identified using the coiled coil directed approach were identified in the comprehensive two-hybrid experiments described above. This result points to a high frequency of false negatives in the comprehensive two-hybrid screens. This observation is similar to that described above where a higher frequency of interactors was found when pairwise two-hybrid tests were performed versus the use of libraries of activation domain clones (Uetz et al., 2000). In addition to the use of a pairwise screen, the coiled coil experiment may have also identified more interactors because only the coiled coil regions of the yeast proteins were used in the screen. The use of full-length proteins in the two-hybrid screen could obscure interactions that are detected only when using fragments of proteins. For example, the full-length protein may contain an interaction site that is masked until an allosteric change in the protein reveals it (Hu, 2000). Taken together, these experiments indicate that estimates of the yeast interactome based solely on two-hybrid assays are underestimates. Thus, computationally directed screens that can approach saturation, such as the coiled coil motif screen, are likely to provide information not obtained in the comprehensive library screens.

Network of protein-protein interactions in yeast

The data from the large-scale two-hybrid analyses of protein-protein

interactions in yeast has been combined with interactions detected by other biochemical methods and reported in the scientific literature to gain a more comprehensive view of interactions in the entire proteome. Schwikowki et al. (2000) analyzed 2,709 published interactions involving 2,039 yeast proteins, available from public databases and from the large-scale two-hybrid experiments. Surprisingly, they identified a single large network of 2,358 interactions among 1,548 proteins as well as several smaller networks (Schwikowski et al., 2000). In addition, a software program was developed to visualize the interaction network (Fig. 5.4)

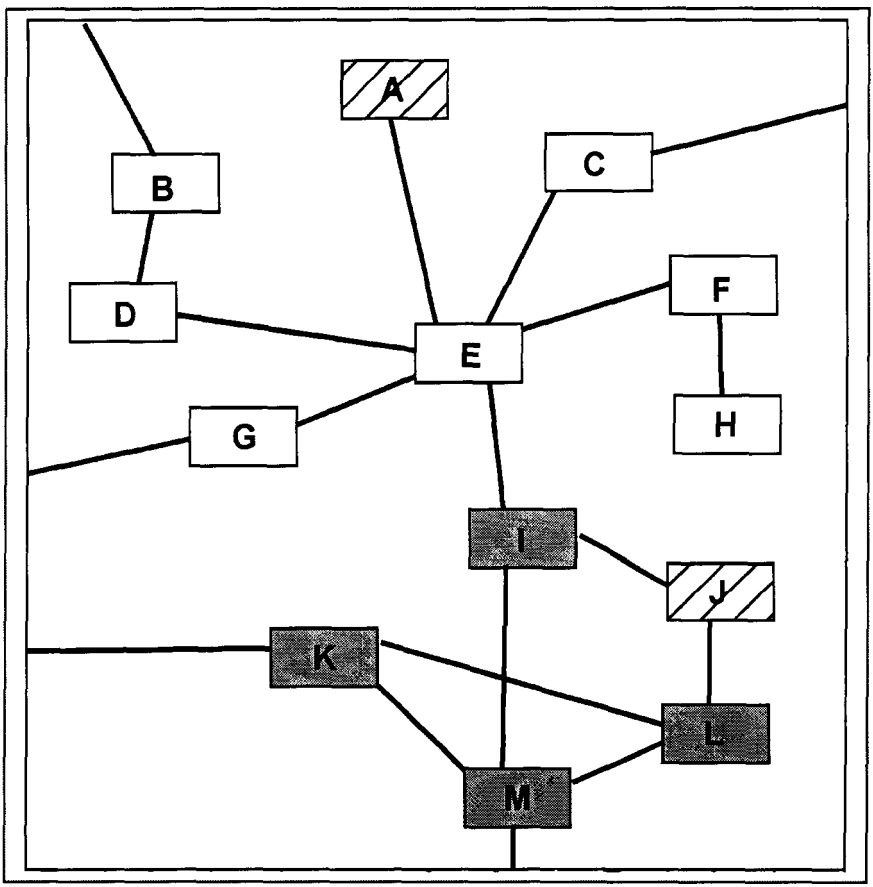


Figure 5.4. Example of a small region of a hypothetical protein interaction network. Each letter represents a different yeast protein. The white boxes and gray boxes represent genes that are involved in the same function while the hatched boxes indicate proteins of unknown function. The A protein is likely to be involved in the same process as the white box protein and the J protein is likely to be involved in the same process as the gray box proteins because of the multiple interactions within the network. The connection between the E and I proteins indicates communication between the cellular processes. Figure adapted from Hazbun and Fields (2001).

There are several interesting features of the large network of interacting proteins. First, proteins of similar function tend to cluster together within the network. For example, 89% of all annotated proteins involved in chromatin structure are located in clusters with other proteins involved in chromatin structure (Schwikowski et al., 2000). In total, 63% of all interactions occur between proteins with a common functional assignment. Second, proteins with a common subcellular location tend to cluster together within the network (Schwikowski et al., 2000). Third, the network of interactions reveals sets of interactions that link cellular processes and uncover crosstalk between cellular compartments. For example, cell cycle control proteins exhibited the most cross-connections between cellular processes. Interactions were detected between proteins in the cell cycle control cluster and proteins involved in mitosis, protein degradation, mating response, DNA synthesis, transcription, signal transduction as well as other processes (Schwikowski et al., 2000). These cross-connections reflect the central role of the cell cycle control proteins in regulating other cellular processes. Finally, the network can be used to form hypotheses about the function of unknown proteins. The network approach to function prediction involves identifying the most common function among the interaction partners of a protein of unknown function and assuming the protein of interest shares the same or a related function (Mayer and Hieter, 2000; Schwikowski et al., 2000).

A current limitation of assigning protein function based on interaction partners is the relative lack of knowledge about the function of proteins within a genome. For example, of the 554 proteins of unknown function within the large network identified by Schwikowki et al. (2000), only 69 had two or more partners of known function. As knowledge of the function of proteins within the genome improves, this approach will become much more powerful.

Architecture of protein networks

The organization of a large network of protein interactions in yeast has been studied in detail (Jeong et al., 2001). The network that was analyzed consists of 1,870 protein connected by 2,240 physical interactions. One goal of the study was to determine if the architecture of the network is best described by a uniform exponential topology, with proteins on average possessing the same number of links to other proteins, or by a heterogeneous scale-free topology, in which proteins exhibit widely different connectivities. The analysis of probabilities of interactions indicated a highly heterogeneous scale-free network in which a few highly connected proteins play a critical role in mediating interactions among a large number

of less connected proteins (Jeong et al., 2001). This type of network architecture, which follows a power-law distribution, is common to other complex systems including the Internet and metabolic networks (Barabasi and Albert, 1999; Jeong et al., 2000)

The heterogeneous architecture of the network suggests it is tolerant to random mutations in that the majority of these lesions would occur in proteins that are not highly connected to other proteins (Jeong et al., 2001). The network, however, is predicted to be highly vulnerable to mutations at the multiply connected node positions. These ideas were tested by rank-ordering all interacting proteins based on the number of links they exhibit, and correlating this with the phenotypic effect of a deletion of the corresponding gene from the genome (Jeong et al., 2001). The correlation was aided by the large data set available from systematic gene disruption experiments performed on the yeast genome (Ross-Macdonald et al., 1999; Winzeler et al., 1999). It was found that the likelihood that removal of a protein will be lethal to the cell correlates with the number of interactions the protein exhibits (Jeong et al., 2001). For example, proteins with five or fewer interactions constitute 93% of the yeast proteins for which gene disruption data is available and yet only 21% of these are essential. In contrast, only 0.7% of the yeast proteins exhibit more than 15 links, but deletion of 62% of these proves lethal (Jeong et al., 2001). Thus, highly connected proteins that are central to the architecture of the network are much more likely to be essential than proteins that have few connections.

An analysis of the architecture of the protein interaction network of the bacterium *Helicobacter pylori* yielded similar results to that of the yeast network (Jeong et al., 2001; Rain et al., 2001). It has therefore been predicted that systematic protein-protein interaction studies of other organisms will reveal networks with similar architecture (Jeong et al., 2001). Interesting questions remain to be answered about the evolution of protein interaction networks and the properties of proteins situated at highly connected node positions. For example, it has been proposed that proteins at node positions may share common structural features that enable them to bind to many different proteins (Hasty and Collins, 2001). Consistent with this idea is the finding the that hinge region on the Fc fragment of human immunoglobulin G interacts with at least four different natural protein scaffolds and also serves as the binding site for random peptides evolved to bind the Fc fragment (DeLano et al., 2000). Characterization of this consensus-binding site indicated it is an adaptive, exposed, nonpolar, and energetically important region on the surface of Fc that is primed for interaction with a variety of different molecules (DeLano et al., 2000). It

will be of interest to determine if the node proteins contain surfaces with similar intrinsic physiochemical properties that favor protein-protein interactions.

5.3 Genome-wide yeast two-hybrid analysis of other organisms

Two-hybrid analysis protein-protein interactions in viral systems

Comprehensive two-hybrid experiments have also been performed using ORFs from viral, bacterial, and animal genomes (Bartel et al., 1996; McCraith et al., 2000; Rain et al., 2001; Walhout et al., 2000). Among the viral systems, the bacteriophage T7 and vaccinia virus genomes have been examined. The advantage of the viral genomes is that it is possible to systematically test all possible pairwise interactions. For example, the set of 266 predicted ORFs from vacccinia virus were cloned as fusions to the Gal4 activation domain as well as the Gal4 DNA-binding domain (McCraith et al., 2000). Each of the potential 70,756 pairwise combinations of proteins were assayed by mating each DNA binding domain fusion to an array of the 266 activation domain fusions. A total of only 37 protein-protein interactions were identified, of which 28 were previously unknown (McCraith et al., 2000). As the authors state, this is likely to be only a fraction of the interactions that occur during a viral infection. One reason for the low number of interactions could be the large number of vaccinia proteins that are membrane associated. These proteins were expressed as full length ORFs and are unlikely to reach the yeast nucleus to participate in two-hybrid interactions (McCraith et al., 2000). Expression of full-length ORFs may also mask certain interactions and lead to a high frequency of false negatives (Hu, 2000). The important message from these experiments is that, even if a screen is saturating for all pairwise combinations of protein-protein interactions, a significant proportion of interactions will not be identified using the two-hybrid screen exclusively.

Two-hybrid analysis ofprotein-protein interactions in bacteria

The genome of the *Helicobacterpylori* bacterium is 1.6 million base pairs in size and encodes 1590 ORFs (Tomb et al., 1997). The comprehensive two-hybrid library screen performed with these ORFs differs from the yeast experiments described above in that the Gal4 activation domain library used consisted of over ten million random genomic fragments (Rain et al., 2001). Thus, the potential problem of full-size ORFs masking protein-protein interactions is reduced. A total of 261 ORFs were fused to the Gal4 DNA binding domain to create a set of baits. These ORFs

were selected to avoid hydrophobic proteins that may not target to the yeast nucleus (Rain et al., 2001). The activation domain library was mated to each of the 261 DNA binding domain fusions to screen for interactions. A total of 1,200 interactions were identified, which connected nearly 50% of the genome. This approach seems a very efficient means of avoiding the problems associated with ORF by ORF pairwise screens. In addition, sequence data is accumulated from multiple fragments within each protein interactor identified from the activation domain library. Alignment of the fragments therefore allows one to map the region of interaction within the protein (Rain et al., 2001).

Two-hybrid analysis of protein-protein interactions in Caenorhabditis elegans

In contrast to the viral and bacterial systems, the *Caenorhabditis elegans* genome is large, encoding approximately 20,000 predicted ORFs. A comprehensive pairwise screen of all of the ORFs would require 4.0 x 10^8 matings, which is not feasible with present technology. However, a directed screen has been performed using 27 proteins known to be involved in vulval development fused to the Gal4 DNA binding domain and a *C. elegans* cDNA library fused to the Gal4 activation domain (Walhout et al., 2000). The two-hybrid assay identified 148 interactions involving 124 different proteins. In an attempt to determine the biological relevance of the identified interactions, the authors searched for conserved interactions in orthologous proteins from other organisms (Walhout et al., 2000). Such conserved interactions were labeled "interlogs" and their identification is a useful method to increase confidence in the validity of an interaction. In addition, the authors used a systematic clustering analysis to search for networks of interactions that form closed loops; reasoning that ORFs included in such loops have an increased likelihood of being involved in biologically significant interactions (Walhout et al., 2000).

It is apparent from the large amount of data from several organisms that the two-hybrid system is an efficient, automatable and thus high-throughput method for identifying protein-protein interactions on a genomewide basis. However, it is also apparent that a significant proportion of false positives and false negatives are inherent to the assay. Therefore, other methods of identifying protein-protein interactions are required to both validate two-hybrid data and to discover new interactions.

5.4 Bacterial two-hybrid system for the detection of protein-protein interactions

Two-hybrid system based on activation of E. coli RNA polymerase

A bacterial two-hybrid system has been developed that, similar to the yeast system, functions via activation of transcription (Dove and Hochschild, 1998; Joung et al., 2000). RNA polymerase (RNAP) in *E. coli* consists of an enzymatic core composed of the α, β, and β' subunits in the stoichiometry $\alpha_2\beta\beta'$, and one of several σ factors that enable the enzyme to recognize specific promoters (Hellman and Chamberlin, 1988). Many bacterial transcriptional activator proteins bind the promoters they regulate and interact directly with subunits of RNAP. The most commonly observed contact is between activator proteins and the α subunit of RNAP (Ebright and Busby, 1995). The function of the α subunit is to initiate the assembly of RNAP by forming a dimer (Igarashi et al., 1991).

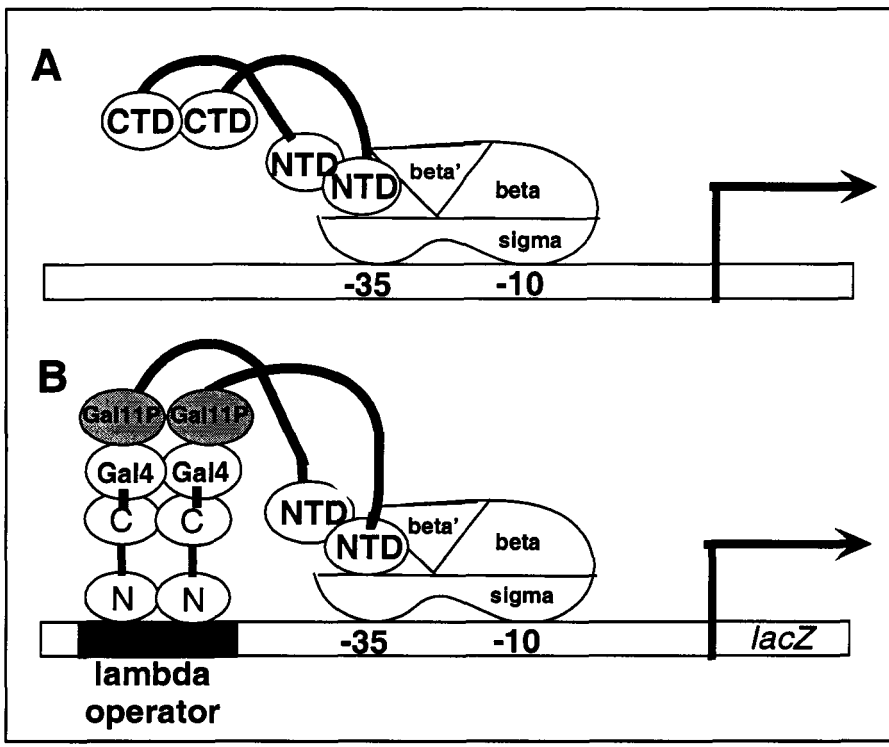


Figure 5.5. **A.** Schematic illustration of the *E. coli* RNA polymerase showing the domain structure of the α subunit. The α-NTD domain is responsible for assembly of RNAP while the α-CTD domain binds DNA and is a target for transcriptional activators. **B.** The two-hybrid system is based on the interaction of proteins that are fused to the λ repressor and NTD domain of the α subunit of RNAP. In the example shown, Gal4 interacts with Gal11P to recruit RNAP to the promoter and activate transcription of the *lacZ* reporter gene. Figure adapted from Dove and Hochschild (1998).

The bacterial two-hybrid system takes advantage of the domain structure of the α subunit of RNAP. The α-NTD domain is responsible for the assembly reaction of RNAP and the α-CTD domain, which is connected to the α-NTD by a flexible linker region, can bind DNA and is the target of several transcriptional activator proteins (Ebright and Busby, 1995). Two-hybrid activation was demonstrated by replacing the α-CTD domain with a section of the Gall 1P protein. The second component of the system consists of the Gal4 protein fused to the C-terminus of the DNA-binding λ-repressor protein (Dove and Hochschild, 1998). The λ-repressor-Gal4 fusion protein binds to the λ-operator sequence, which is placed next to the *lacZ* gene encoding the reporter enzyme, β-galactosidase (Fig. 5.5). The Gal4 protein interacts with the $Gal11^P$ domain of the α-NTD-$Gal11^P$ fusion, which, in turn, recruits RNAP to the promoter and actives transcription of the *lacZ* reporter gene (Dove and Hochschild, 1998).

A bacterial one-hybrid system for the detection of protein-DNA interactions has also been developed (Joung et al., 2000). This system is similar to the two-hybrid system described above except the λ-repressor DNA-binding protein was replaced with the zinc finger DNA-binding protein, Zif268 (Joung et al., 2000). In addition, the reporter screen was converted to a selection by replacement of the *lacZ* reporter with the yeast HIS3 gene. HIS3 encodes an enzyme required for histidine biosynthesis that can complement the growth defect of *E. coli* cells bearing a deletion in the homologous *hisB* gene (Joung et al., 2000). Use of this reporter system allowed for the selection of Zif268 mutants with altered DNA-binding properties from libraries containing >108 unique clones.

The bacterial one and two-hybrid systems have potential advantages over the yeast two-hybrid system due to the higher transformation efficiency and faster growth rate of *E. coli.* To date, however, the bacterial two-hybrid system has not been used for genome-scale analysis of protein-protein interactions.

5.5 Use of phage display to detect protein-ligand interactions

Display of proteins on M13 filamentous phage

Phage display is a powerful technique for studying protein-ligand interactions (reviewed by (Smith and Petrenko, 1997)). The most common implementation of the method involves the fusion of peptides or proteins to a coat protein of a filamentous bacteriophage (Smith, 1985). The peptides or proteins are normally fused to the N-terminus of either the gene *III* or gene *VIII* phage proteins. The gene *III* protein is a minor coat (3-5 copies per phage) protein located at the tip of the phage and is responsible for

attachment of the phage to the bacterial F pilus in the course of the normal infection process (Riechmann and Holliger, 1997). The gene VIII protein is the major coat protein that is present in 2700 copies per phage particle (Rasched and Oberer, 1986). Because the gene encoding the fusion protein is packaged within the same phage particle, there is a direct link between the phenotype, i.e., the ligand binding characteristics of a displayed protein, and the DNA sequence of the gene for the displayed protein. This permits large libraries of peptides of random amino acid sequence to be rapidly screened for desired ligand binding properties (Fig. 5.6) (Smith, 1985). In addition, large collections of mutants of a displayed protein can be screened for variants with altered ligand binding characteristics (Katz, 1997).

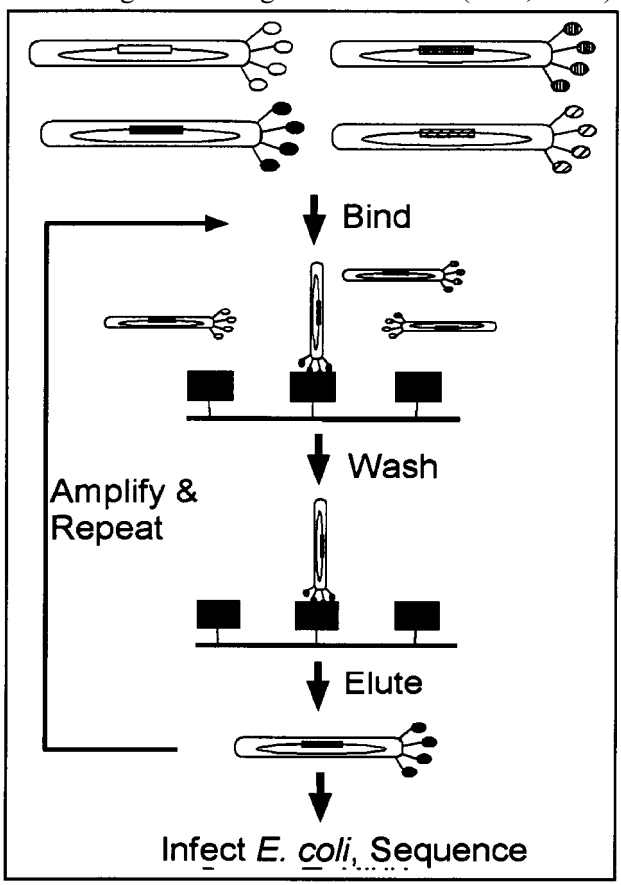


Figure 5.6. Filamentous phage display. Random sequence peptides or variants of proteins are fused to the gene III capsid protein of the M13 bacteriophage. The target ligand is immobilized in the solid phase. Phage displaying a protein that interacts with the target are enriched by affinity purification in a process called panning. After multiple rounds of panning, the phage are used to infect E. coli and the identity of the selected inserts is determined by DNA sequencing.

Jun-Fos phage display system

More recently, Jacobson and Frykberg have extended the shotgun cloning of fragmented genes to encompass entire bacterial genomes (Jacobsson and Frykberg, 1995; Jacobson and Frykberg, 1996; Jacobsson et al., 1997). In test experiments, total genomic DNA of *Staphylococcus aureus* was fragmented by sonication and cloned into a gene *III* and a gene *VIII* phage display vector. The potential of the method was demonstrated by selecting fragments of the gene encoding protein A from the *S. aureus* genomic library by enriching for phage that bind to immobilized IgG protein (Jacobsson and Frykberg, 1995; Jacobsson and Frykberg, 1996). This system has also been used to clone a fibrinogen binding protein from *Staphylococcus epidermidis* (Nilsson et al., 1998) as well as surface proteins from group C streptococci that bind alpha (2)-macroglobulin, serum albumin and IgG (Jacobsson et al., 1997). A limitation of this approach, however, is that the random fragments of DNA must be cloned between a signal sequence and the N-terminus of gene *III* encoding protein (g3p) or gene *VIII* encoding protein (g8p). Therefore, only one in 18 clones (3 x 3 x 2) will contain a fusion that is in the correct orientation and is in-frame with both the signal sequence and the phage coat protein. The standard phage display system is also not suited to the construction of cDNA libraries from eukaryotic organisms because the presence of the stop codon at the end of the gene encoded in cDNA precludes fusion to the phage capsid protein.

To partially overcome the problem of most inserts being out of frame and to permit the cloning of cDNAs, a modified phage display vector that takes advantage of the high affinity protein-protein interactions of the Jun and Fos leucine zipper proteins has been developed (Crameri and Suter, 1993). For this system, DNA encoding the Jun leucine zipper domain is expressed from a *lac* promoter as a fusion protein with the phage coat protein g3p, which results in the Jun leucine zipper being displayed on the surface of the phage particle (Fig. 5.7). On the same plasmid, DNA encoding the *fos* leucine zipper is fused to a bacterial signal sequence and is also expressed from a *lac* promoter. Restriction enzyme cloning sites are present at the 3'-end of the *fos* gene to insert cDNA libraries (Crameri and Suter, 1993; Crameri et al., 1994). During the course of phage propagation the Jun-geneIII protein is secreted to the periplasm of *E. coli* where it is incorporated into the phage particle (Fig. 5.7). The Fos fusion to the protein of interest is also secreted to the periplasmic space of *E. coli* where it interacts with the Jun-geneIII protein via the leucine zipper interaction. Disulfide bonds are engineered into the Jun and Fos leucine zippers to result in a covalent linkage between the protein of interest and the phage particle via the Jun-Fos interaction (Crameri and Suter, 1993) (Fig. 5.7).

The advantage of the Jun-Fos phage display system is that the fusion of the library DNA is to the C-terminus of a protein rather than between the

signal sequence and the N-terminus. Therefore, one in six (3 x 2) inserts will fuse in the correct reading frame and orientation. In addition, for shotgun cloning of genomic fragments, the size of the insert is less critical because the presence of a naturally occurring stop codon at the end of an ORF does not affect expression of the fusion as it does with the direct fusions to gene *III* or gene *VIII.* The pJuFo system has been used to fuse the expression products of a cDNA library from *Aspergillus fumigatus* on the surface of phages and screen for phages that bind human serum IgE to identify and clone allergenic proteins produced from *A. fumigatus* (Crameri et al., 1994). In addition, a phage cDNA expression library from human lymphocytes has been used to identify fl-actin as a cellular protein that binds to HIV-1 reverse transcriptase (Hottiger et al., 1995). Finally, a genomic library of *E. coli* has been constructed in pJuFo and used to clone genes based on binding affinity to a target (Palzkill et al., 1998).

The pJuFo system also has the potential to display homodimers or homomultimeric proteins on the surface of the phage. This is possible because the protein of interest is expressed independent of the phage proteins. Therefore, the Fos-ORF fusion protein is secreted to the periplasm, the Fos portion interacts with Jun to attach the protein to the phage, and other copies of the Fos-ORF protein can interact with each other to form a dimer or multimer. However, the successful display of homodimers has not been demonstrated in practice.

The pJuFo approach circumvents the problem of most inserts being out of frame; however, another potential problem associated with display of proteins on filamentous phage is whether the protein of interest can be efficiently secreted. Filamentous phage does not lyse *E. coli* during propagation; instead the phage particles are extruded out of living cells (Russell, 1991). Because of this property, gene *III* and gene *VIII* fusion proteins are secreted to the periplasm for viral assembly. Therefore, if a protein is not naturally secreted it may not be efficiently packaged into a phage particle. It is unclear to what extent this limits phage display of ORF encoded proteins but it is known that many proteins are not displayed efficiently (Cochrane et al., 2000). This may explain why filamentous phage display has been an effective means of cloning antibodies but has not been used extensively to map protein-protein interactions using cDNA libraries.

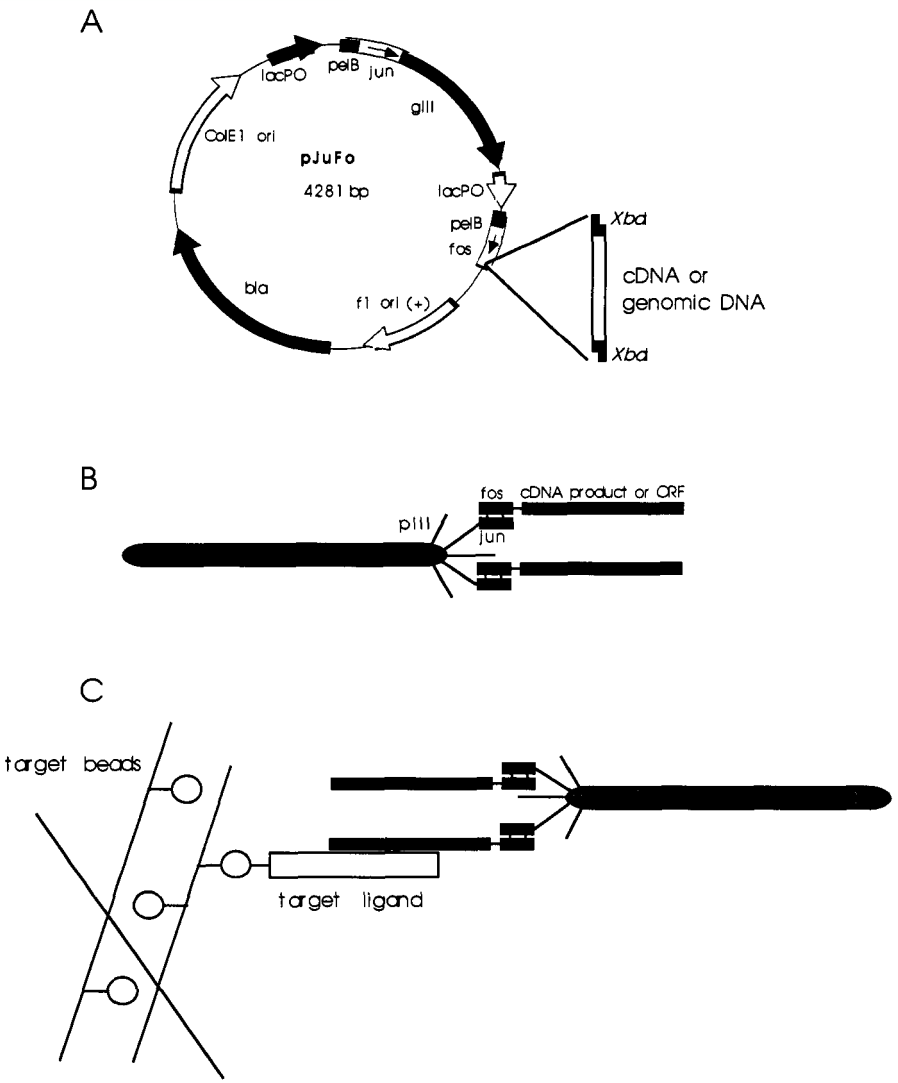


Figure 5.7. Jun-Fos filamentous phage display system. **A.** cDNA or fragmented genomic DNA is inserted as a fusion to the C-terminus of Fos. **B.** Phage particles produced from *E. coli* containing the library display the cDNA products or ORFs via linkage between Jun and Fos. **C.** Phage amplification and cloning of a specific insert is due to binding of the phage containing the plasmid encoding a fusion protein that is able to bind a target ligand immobilized on beads or microtiter plates. The majority of the phages from the expression library do not encode a protein that can bind the ligand and are therefore washed off the beads. The binding phages are released from the beads, amplified in *E. coli,* and the process is repeated to enrich for phages that preferentially bind the target. Finally, the phages are used to infect *E. coli* and the identity of inserts from individual clones is determined by DNA sequencing.

Display ofproteins on the T7 bacteriophage

The recent development of phage display systems that use lytic bacteriophage vectors such as lambda (Maruyama et al., 1994; Santini et al., 1998), T4 (Ren et al., 1996), and T7 (Rosenberg et al., 1996) has provided an alternative that is independent of the *E. coli* secretion machinery. All of these systems have the additional advantage that cloned proteins are fused to the C termini of phage capsid proteins, which facilitates genomic and cDNA library constructions (Fig. 5.8). The display of cDNA libraries on T7 phage has recently been used to identify interactions among signaling proteins in the EGF-receptor signaling pathway (Zozulya et al., 1999). In addition, the T7 system has been used to identify protein-small molecule interactions using a cDNA library (Fig. 5.8) (Sche et al., 1999). These studies suggest phage display can be used for high-throughput protein-protein interactions studies.

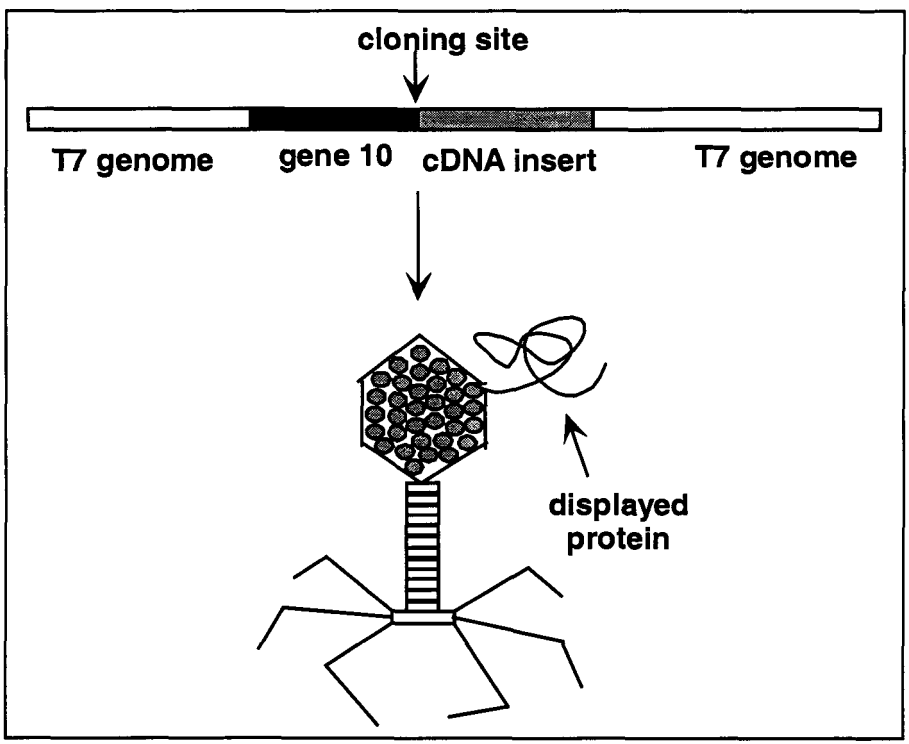


Figure 5.8. Schematic illustration of T7 bacteriophage display system. The gene encoding the protein of interest is fused to the gene 10 protein that encodes a coat protein in the head of the phage. Cloning is to the C-terminus of coat protein 10. Figure adapted from Sche et al. (1999).

An advantage of the genomic or cDNA phage display libraries, be they on filamentous or lytic phages, is that the same library can be used to isolate binding partners for many different ligands. In addition, the ligand used for screening is not restricted to other proteins in that any molecule which can be immobilized can be used as a target for panning. This enables the library to be searched for proteins that bind DNA, RNA, proteins, carbohydrates, amino acids, or other small molecules. Peptide phage display libraries have even been used to probe the vasculature of living animals (Pasqualini and Ruoslahti, 1996). These studies are performed by injecting a phage library into the circulation of a mouse, waiting a short time, harvesting the organ or tissue of interest, and using a homogenate of the tissue to infect *E. coli* in order to amplify the phage that bound to this tissue (Pasqualini and Ruoslahti, 1996). This approach has been used to identify peptides that home specifically to the vasculature of many different tissues and of tumors (Pasqualini et al., 1997; Rajotte et al., 1998). These results suggest the vasculature of most tissues displays markers selective for that tissue. It would be of interest to perform *in vivo* phage display experiments using genomic or cDNA libraries from pathogenic microbes to examine tissue targeting mediated by microbial ORFs.

The ability to identify binding proteins for many types of ligands will make phage display a useful tool for proteomics. Genome sequencing has identified many open reading frames for which no function has been assigned. Genomic phage display libraries could be used to classify open reading frames by binding function. For example, genomic DNA could be immobilized on beads and the set of DNA binding proteins could be selected from the genomic phage display library. By using different classes of ligands, it should be possible to build large categories of open reading frames based on binding properties.

5.6 Detecting interactions by protein fragment complementation assays

Overview

Protein fragment complementation assays are based on an enzyme reassembly strategy whereby a protein-protein interaction promotes the efficient refolding and complementation of enzyme fragments to restore an active enzyme. The approach was initially developed using the reconstitution of ubiquitin as a sensor for protein-protein interactions (Johnsson and Varshavsky, 1994). Ubiquitin is a 76 amino acid protein that

is present in cells either free or covalently linked to other proteins. Ubiquitin fusions are rapidly cleaved by ubiquitin-specific proteases, which recognize the folded conformation of ubiquitin. If ubiquitin is expressed as a fusion to a reporter protein, the cleavage reaction can be followed *in vivo* by release of the reporter (Johnsson and Varshavsky, 1994). However, if a C-terminal fragment of ubiquitin is expressed as a fusion to a reporter and the N-terminal fragment of ubiquitin is expressed separately in the same cell, cleavage does not occur. If, on the other hand, proteins that interact are fused to the N- and C-terminal fragments of ubiquitin, the full-sized ubiquitin is reconstituted and cleavage occurs (Fig. 5.9) (Johnsson and Varshavsky, 1994). Therefore, proteins interactions can be tested *in vivo* by fusing them to the N- and C-terminal fragments of ubiquitin and assaying for release of the reporter protein (Fig. 5.9). As discussed below, the approach has been extended to other systems and now represents a viable alternative to two-hybrid methods for detecting protein-protein interactions *in vivo.*

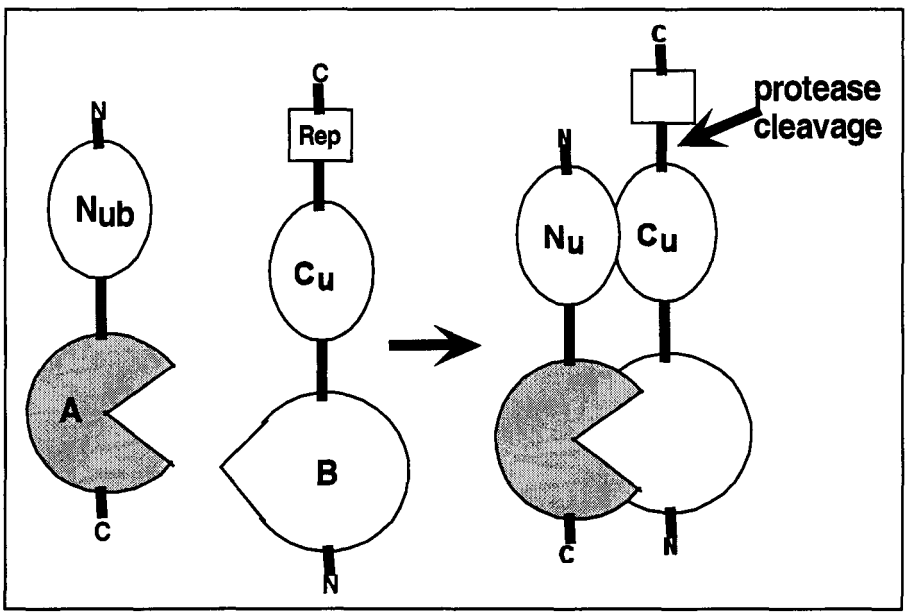


Figure 5.9. Split ubiquitin as a sensor for protein-protein interactions. Protein A is fused to the N-terminal domain and protein B is fused to the C-terminal domain of ubiquitin. Interaction of A and B reconstitutes a full-sized, folded ubiqutin. The folded ubiquitin is recognized by a specific protease and cleavage releases the reporter protein.

Protein fragment complementation using dihydrofolate reductase

The dihydrofolate reductase enzyme (DHFR) is involved in one-carbon metabolism and is required for the survival of prokaryotic and eukaryotic cells. The enzyme catalyzes the reduction of dihydrofolate to tetrahydrofolate, which is required for the biosynthesis of serine, methionine, purines, and thymidylate. The mouse dihydrofolate reductase (mDHFR) is a small (21 kD), monomeric enzyme that is highly homologous to the *E. coli* enzyme (29% identify) (Pelletier et al., 1998). The three-dimensional structure of DHFR indicates that it is comprised of three structural fragments: F[1], F[2] and F[3] (Gegg et al., 1997).

E. coli DHFR is selectively inhibited by the antibiotic trimethoprim. The mDHFR enzyme, however, has 12,000-fold lower affinity for trimethoprim than does bacterial DHFR. Thus, *E. coli* cells expressing mDHFR are able to grow in the presence of trimethoprim. It has been demonstrated that expression of the F[1,2] and F[3] domains of mDHFR separately as fusions to oligomerizing leucine zipper sequences results in trimethoprim resistant *E. coli* cells due reconstitution of the enzyme *in vivo* (Fig. 5.10) (Pelletier et al., 1998). The reconstitution is due to the interaction of the leucine zipper sequences which brings the F[1,2] and F[3] fragments of mDHFR into close proximity where they are able to fold into a single, functional enzyme. Potential protein-protein interactions can therefore be tested by fusing one of the proteins to the F[1,2] domain and its putative binding partner to the F[3] domain and assaying for trimethoprim resistance in *E. coli* cells (Fig. 5.10). This system is analogous to the ubiquitin system described above.

The use of the mDHFR protein fragment complementation assay has recently been extended to mammalian cells (Remy and Michnick, 1999). Reconstitution of mDHFR activity in mammalian cell culture is monitored in DHFR-negative cells grown in the absence of nucleotides. An active mDHFR enzyme is required for the growth of these cells because DHFR activity is required for the biosynthesis of purines and thymidylate (Remy and Michnick, 1999). A second approach for monitoring mDHFR reconstitution in cell culture involves a fluorescence assay based on the detection of fluorescein-methotrexate (fMTX) binding to reconstituted mDHFR *in vivo.* The basis for this assay is that when mDHFR is reassembled in cells it binds with high affinity to fMTX in a 1:1 complex (Remy and Michnick, 1999). Bound fMTX is retained while unbound fMTX is rapidly transported out of the cells. Therefore, fluorescence microscopy, FACS, or spectroscopy can be used to monitor the presence of reconstituted

mDHFR (Remy and Michnick, 1999).

The mDHFR protein complementation assay has been used to map a signal transduction network that controls the initiation of translation in eukaryotes (Remy and Michnick, 2001). A total of 35 different pairs of full-length proteins were analyzed and 14 interactions were identified using the survival selection of cells grown in the absence of nucleotides. In addition, the use of the fMTX reagent in combination with fluorescence microscopy was used to localize the protein complex within cells (Remy and Michnick, 2001).

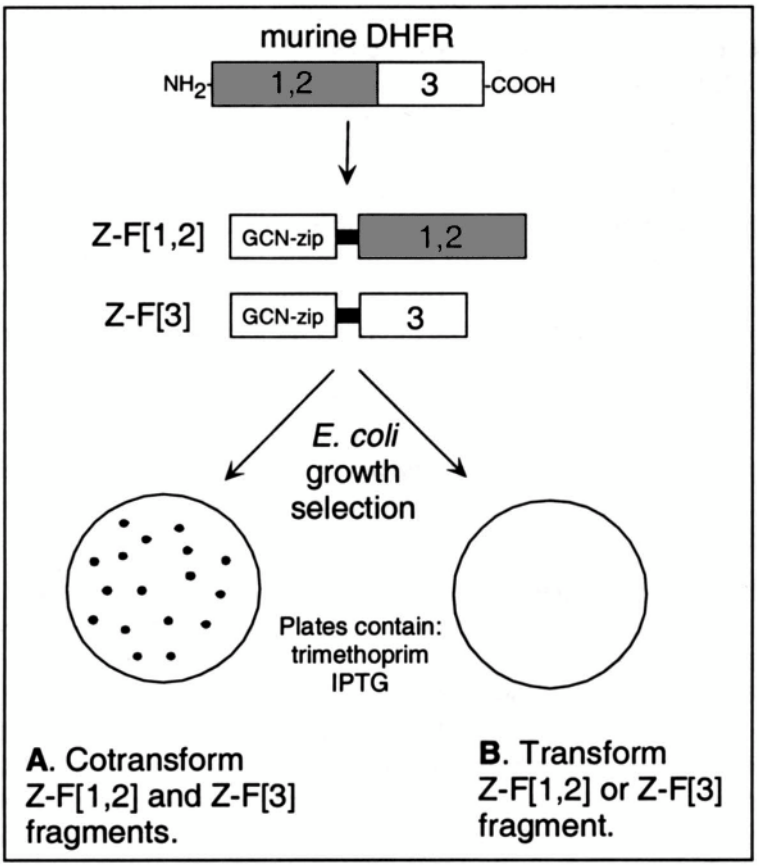


Figure 5.10. Protein complementation assay using murine DHFR. The F[1,2] and F[3] fragments are each fused to the homodimerizing GCN4 leucine zipper protein. A. Transformation of both Z-F[1,2] and Z-F[3] constructs results in reconstituted DHFR and growth of E. coli on agar plates containing trimethoprim. B. Transformation of Z-F[1,2] or Z-F[3] alone does not result in trimethoprim resistant *E. coli* cells. Figure adapted from Pelletier et al. (1998).

The mDHFR protein complementation system possesses potential advantages compared to the yeast two-hybrid system for the detection of protein-protein interactions. For example, the mDHFR system can be used in *E. coli* or in mammalian cells (Pelletier et al., 1998; Remy and Michnick, 1999). The use of *E. coli* may be useful for high-throughput screening because of the ease of manipulation and the fast growth rate. However, when assaying interactions of mammalian proteins it may be advantageous to work with a native system. In addition, the ability to localize interactions within cells using the fMTX reagent and fluorescence microscopy provides important information on the *in vivo* location of protein complexes that is not available from the yeast two-hybrid system. Therefore, it is likely that this technology will find wide use for genome-scale proteomic studies in the future.

Monitoring protein interactions by intracistronic fl-galactosidase complementation

The fl-galactosidase enzyme is widely used as a reporter of gene expression because of its ability to cleave the chromogenic substrate, 5-bromo-4-chloro-3-indoyl β-D-galactopyranoside (X-Gal) to yield a blue product. protein-protein interaction assay has been developed based on the classical bacterial genetic phenomenon of intracistronic complementation (Mohler and Blau, 1996). In *E. coli,* deletions of either the N or C terminus of β-galactosidase produce enzyme that is inactive but that can be complemented by coexpression of a second deletion mutant that contains the regions that are missing from the first mutant. Complementation occurs by assembly of the deleted fragments into a stable octameric protein that contains all of the essential domains of the wild-type homotetramer (Mohler and Blau, 1996). The N- and C-terminal domains present in the mutants involved in complementation are known as the α and ω regions, respectively. The interaction assay is based on the fact that when fl-galactosidase fragments lacking the α domain (Δα) and the ω domain (Δω) are coexpressed, complementation to create an active enzyme is very inefficient (Mohler and Blau, 1996). When the Δα and Δω β-galactosidase mutants are fused to proteins that interact, however, the association of the Δα and Δω fragments is favored and efficient complementation occurs (Mohler and Blau, 1996; Rossi et al., 2000). Therefore, potential protein-protein interactions can be tested by fusing the proteins of interest to the Δα and Δω β-galactosidase deletion mutants and assaying for enzyme function using X-Gal (Mohler and Blau, 1996; Rossi et al., 2000).

The fl-galactosidase complementation assay has also been adapted for use in mammalian cells (Rossi et al., 1997). The availability of fluorescent substrates for β-galactosidase allows for fluorescence microscopy and FACS analysis of mammalian cells expressing the fusion proteins of interest. Therefore, similar to the mDHFR system, fl-galactosidase complementation assays may prove useful for genome-scale studies of protein-protein interactions in mammalian cells.

5.7 Use of mass spectrometry for protein-protein interaction mapping

Overview

As described in Chapter 2, rapid protein identification can be achieved by searching protein and nucleic acid databases directly with peptide mass data generated by mass spectrometry (Yates III, 2000). The most common application of mass spectrometry to protein-protein interaction mapping has involved identifying the components of protein complexes. For these experiments, entire multi-protein complexes are isolated from cells using affinity-based methods (Fig. 5.11). This usually requires knowledge of the identity of at least one protein in the complex. This protein can be tagged with an affinity handle such as glutathione-S-transferase, a poly-histidine repeat or an epitope for an antibody. The tagged protein can then be overexpressed in cells and affinity purified under non-denaturing conditions such that the interaction partners co-purify. The complex is then eluted and individual proteins are resolved by SDS-PAGE; the bands are cut from the gel, proteolyzed with trypsin and the exact mass of the peptides is determined by mass spectrometry. Database searching with the peptide mass data yields the identity of proteins from the complex (Pandey and Mann, 2000; Yates III, 2000). There are numerous examples in the literature illustrating the general approach; including studies of the nuclear pore complex (Rout et al., 2000), the yeast Arp2/3 complex (Winter et al., 1997), TATA-binding-protein-associated factors (Grant et al., 1998), the yeast spindle-pole body complex (Wigge et al., 1998), spliceosome components (Neubauer et al., 1998) and proteins bound to the chaperonin GroEL (Houry et al., 1999).

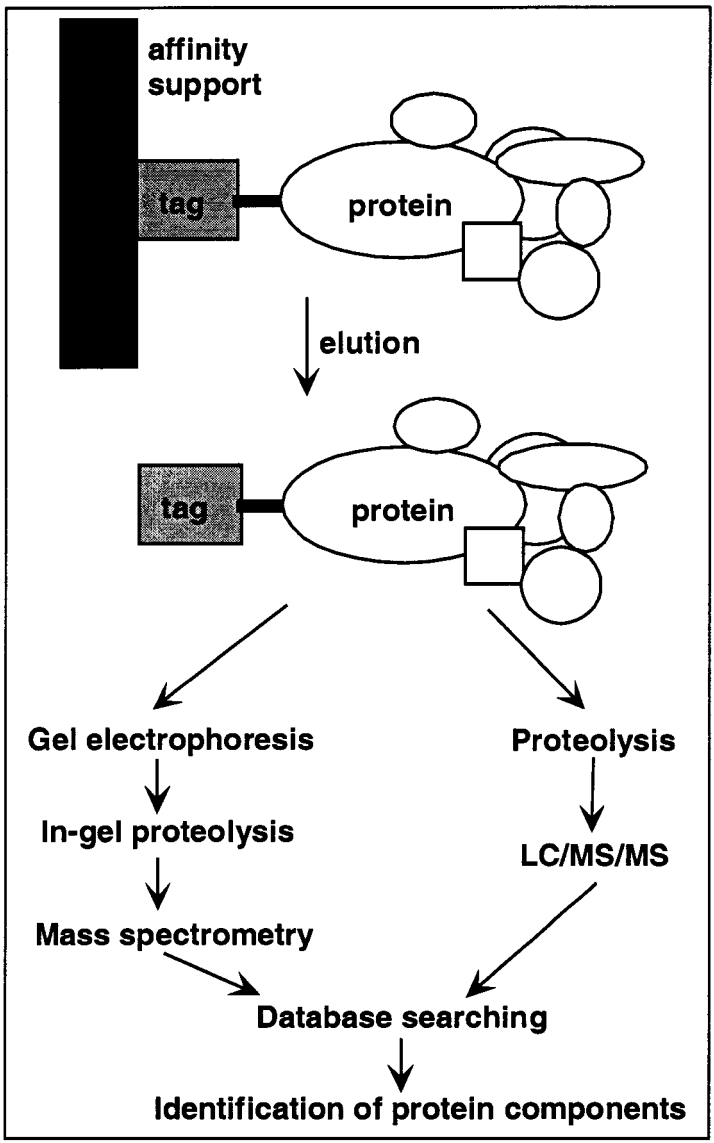


Figure 5.11. Generic approaches to identify interacting proteins within complexes. The complex is isolated from cells by affinity purification using a tag sequence attached to a protein known to be in the complex. Alternatively, the complex can be immunprecipitated with an antibody to one of the proteins in the complex. The proteins are resolved by polyacrylamide gel electrophoresis, proteolyzed, and the mass of the resulting peptides is determined by mass spectrometry. Alternatively, the proteins can be proteolyzed and the resulting peptides resolved by liquid chromatography. The peptide masses are then determined by mass spectrometry and used for database searching to identify the component proteins.

Example: Identification of substrates for E. coli GroEL

The general approach of using protein purification coupled with mass spectrometry to identify protein interaction complexes will be illustrated by a study on the bacterial GroEL protein (Houry et al., 1999). The chaperonin GroEL, along with its cofactor GroES, is essential for growth of *E. coli* (Fayet et al., 1989). The function of GroEL is to facilitate folding by limiting the side reaction of aggregation. GroEL is a homo-oligomer of 14 subunits that forms a cylindrical structure with two large cavities. Substrate protein binds in the central cavity of the cylinder via hydrophobic surfaces exposed within the GroEL complex. GroES then binds the apical surface of the cylinder to trap the substrate in an enclosed cavity where aggregation is prevented. Approximately 10% of new translated peptides are known to interact with GroEL but the identity of these substrate proteins was not known.

The preferred substrates for GroEL were identified by pulse-chase labeling of growing *E. coli* cells (Houry et al., 1999). At various times of chase, the GroEL-substrate complexes were isolated by immunoprecipitation with anti-GroEL antibodies. The precipitated proteins were resolved on two-dimensional polyacrylamide gels and their identity was determined by digestion of the spots with trypsin followed by peptide-mass fingerprint analysis using MALDI-TOF mass spectrometry (Houry et al., 1999). In this way, a total of fifty-two different proteins were identified as substrates for GroEL. These proteins were analyzed for a common structural motif that may form the basis of their interaction with GroEL. It was found that GroEL substrates preferentially contain several $\alpha\beta$ domains compared to other *E. coli* proteins (Houry et al., 1999). These experiments illustrate the power of the mass spectrometry approach to identifying protein-protein interactions in that a large set of proteins that bind relatively non-specifically to a target protein were efficiently identified. The other examples listed above illustrate the utility of this method for efficiently identifying the components of large structures such as the nuclear pore. The major limitation of this approach is that the interacting proteins must have sufficient affinity to be retained in the complex during the purification steps. Hence, weakly interacting proteins may be lost from the complex during the purification and therefore may not identified using this approach.

The various methods outlined in this chapter are likely to emerge as important tools for proteomics because protein interaction mapping will dominate proteomic studies over the next several years. Knowledge of interaction partners provides important clues as to biological function and thus protein interaction mapping will be of great value to understanding the biology of the cell.

Chapter 6
PROTEIN-PROTEIN INTERACTION MAPPING: COMPUTATIONAL

The knowledge of protein-protein interaction partners can provide important information on the possible biological function of a protein. As described in the previous chapter, large-scale efforts are underway to detect and analyze protein-protein interactions using experimental methods such as the yeast two-hybrid system are currently underway. However, the ever-increasing amount of genome sequence data makes such mapping efforts a daunting and tedious proposition. Recently, several algorithms have been developed to identify functional interactions between proteins using computational methods. These computational methods are advantageous in that they can provide leads for the experimental methods, which could simplify the task of protein interaction mapping. Also, in contrast to the experimental methods, the predictive capacity of the computational methods improves as more genome sequence data becomes available.

6.1 Computational detection of functional linkages between proteins

Phylogenetic profiles

The phylogenetic profile is a computational method that detects proteins that participate in a common structural complex or metabolic pathway (Pellegrini et al., 1999). The proteins detected are defined as functionally linked. Thus, the proteins identified may physically interact within a complex or may simply be components of a common biochemical pathway. The method takes advantage of the numerous fully sequenced genomes that are now available. The hypothesis supporting the method is that functionally linked proteins evolve in a correlated fashion and, therefore, they have homologs in the same subset of organisms. The idea is that it is very unlikely that two proteins would always both be inherited (or not inherited) to a new species unless they were functionally linked. Thus, if homologs to a pair of proteins are found in the same subset of sequenced

organisms, the proteins are functionally linked (Fig. 6.1).

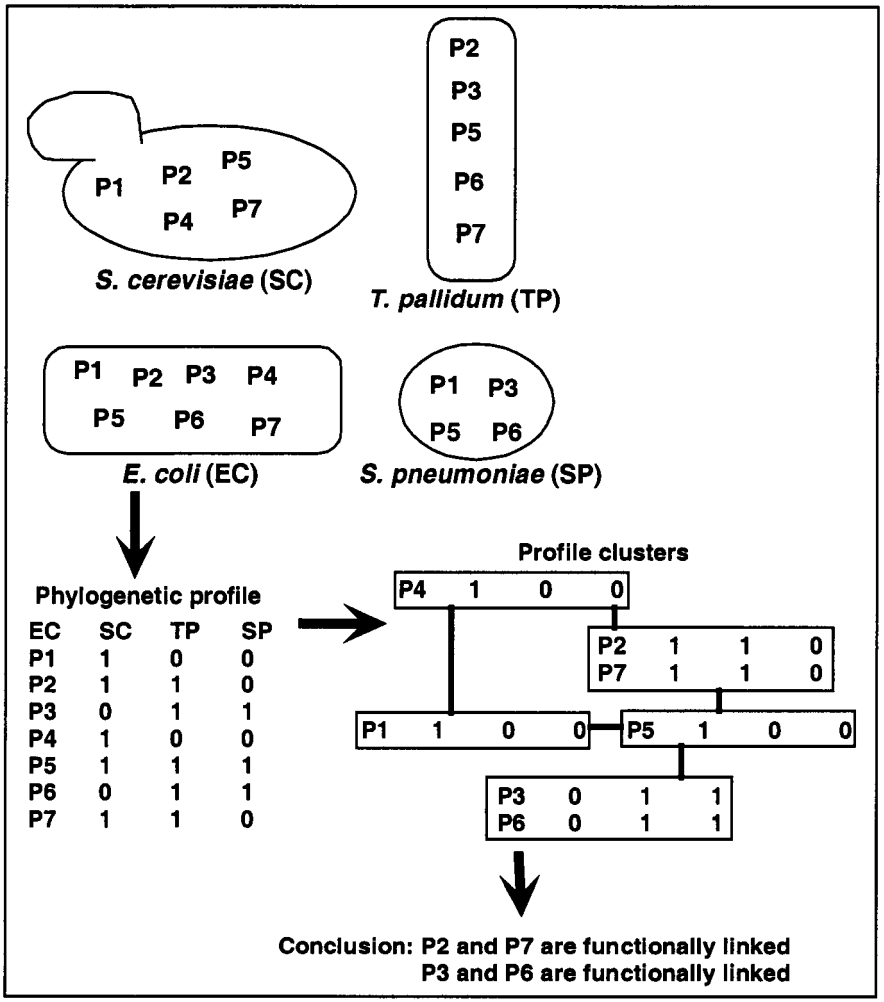


Figure 6.1. Phylogenetic profile method. Four genomes are shown, each containing a subset of proteins (P1..P7). The presence or absence of a protein is given by 1 or 0 in the phylogenetic profile shown at lower left. Identical profiles are shown clustered on the right. In this example, the P2 and P7 are functionally linked because they exhibit the same phylogenetic profile. The P3 and P6 protein are similarly linked. Figure adapted from Pellegrini et al. (1999).

A phylogenetic profile is a description of the presence or absence of a particular protein in a set of organisms whose genome has been sequenced. The profile is a string with n entries, each one bit, where n is the number of

genomes sequenced (Fig. 6.1). The presence of a homolog to a given protein in the *n*th genome is given with an entry of unity while the absence of a homolog is indicated with by an entry of zero (Pellegrini et al., 1999). Proteins are then clustered by the similarity of their phylogenetic profiles. Similar profiles indicate correlated inheritance of proteins and therefore imply a functional linkage between the proteins.

The method was tested by constructing a phylogenetic profile for the 4,290 proteins encoded by the genome of *E. coli* (Pellegrini et al., 1999). The profile was constructed from a total of 16 other fully sequenced genomes. The method accurately predicted functional linkages between sets of ribosomal proteins as well as sets of flagellar proteins. In addition, comparing the keywords from protein database annotations for those proteins predicted to be functionally linked tested functional linkages. The keywords in the protein database give a broad indication of function, if it is known. Therefore, functionally linked proteins would be expected to have similar keyword annotations. A significant correlation was found between keywords for proteins that were predicted to be functionally linked (Pellegrini et al., 1999). Therefore, the phylogenetic profile method appears to be a useful computational tool to provide testable hypotheses about interacting proteins.

Domainfusion or Rosetta Stone method

Functional linkages between proteins have also been detected by analyzing the patterns of fusion of protein domains within sequenced genomes (Enright et al., 1999; Marcotte et al., 1999). This method is based on the observation that some pairs of interacting proteins have homologs in another organism that are fused into a single protein chain. For example, the Gyr A and Gyr B subunits of *E. coli* are separate proteins that are known to interact. Within *S. cerevisiae,* however, the homologs for Gyr A and Gyr B are fused into a single protein, topoisomerase II (Marcotte et al., 1999). Thus, the computational method entails searching through genomic sequences for two proteins, A and B, that, in some other species are expressed as a fused protein, A-B (Fig. 6.2).

The domain fusion method was used for 4290 proteins of the *E. coli* genome to detect functional interactions (Marcotte et al., 1999). A total of 6809 pairs of nonhomologous sequences were found that had significant similarity to a single protein in some other genome. Each of these pairs is a candidate for a pair of interacting proteins in *E. coli.* Comparing the functional annotations for the interacting proteins in the protein database assessed the biological relevance of these interactions. This test is similar to that described above in that, for those cases where protein function is known, each pair of proteins that are presumed to interact should possess a

similar keyword annotation. Of the 3950 *E. coli* pairs of known function, 68% shared at least one keyword in the annotation (Marcotte et al., 1999). This correlation is significantly higher than that found when protein pairs are chosen at random (15%). Therefore, if the function of one protein in a pair is known, the function of the other member can be predicted with reasonable accuracy (Marcotte et al., 1999). Like the phylogenetic profiling method described above, this method will be a useful means of generating hypotheses that can be tested experimentally.

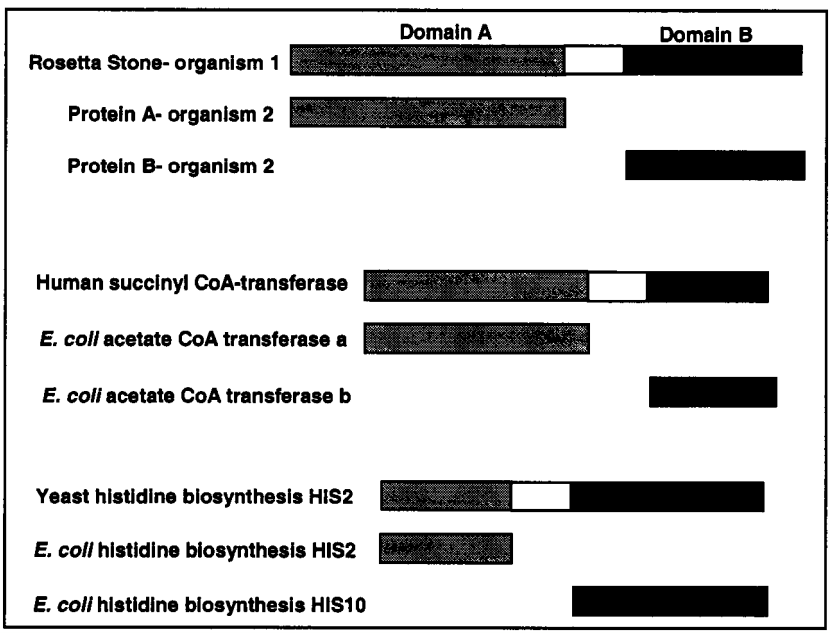


Figure 6.2. Domain fusion method to detect functional linkages between proteins. If two proteins, A and B, from a certain species are expressed as a single fused protein within another species, the proteins A and B are proposed to be functionally linked. Because the fusion protein contains homology to both A and B, it is termed a Rosetta Stone sequence. Specific examples from domain fusion analyses are shown. Figure adapted from Eisenberg et al. (2000) and Marcotte et al. (1999).

Gene neighbor method

A third approach for detecting functional linkages between proteins from genome sequences is the gene neighbor method (Dandekar et al., 1998; Overbeek et al., 1999). The hypothesis underlying this method is that, if on the chromosomes of several genomes, the genes that encode two proteins are neighbors, the proteins encoded by those genes must interact or be involved in a similar function (Fig. 6.3). With the rapid increase in the availability of

genome data, it has been possible to apply this method on a wide scale. Application of the method to a set of prokaryotic genomes lead to a prediction of approximately 100 protein pairs that are hypothesized to interact (Dandekar et al., 1998). For at least 75% of the conserved gene pairs, physical interactions between the encoded proteins have been experimentally demonstrated. These include numerous interactions among ribosomal proteins as well ATP synthase subunits, RNA polymerase subunits, and cell-division proteins (Dandekar et al., 1998). The finding that the majority of protein pairs predicted to interact have been shown experimentally to interact provides strong validation that the method can be used to predict previously unknown interactions.

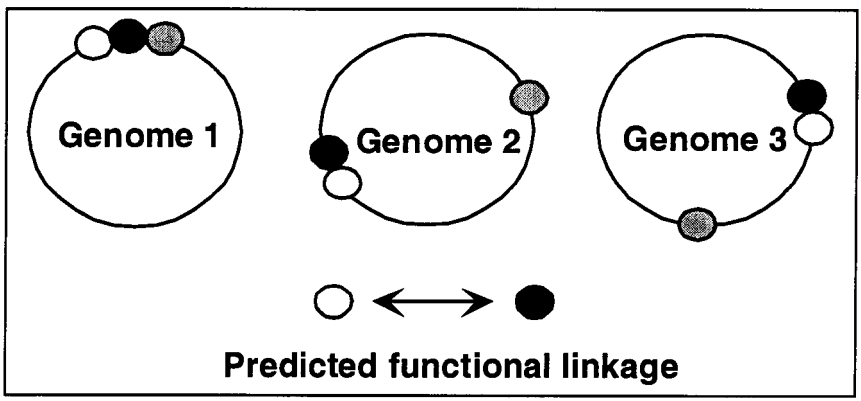


Figure 6.3. Gene neighbor method for detecting functional interactions among proteins. If two genes (white and black circles) are found as neighbors in several genomes, the encoded proteins are predicted to functionally interact. Figure adapted from Eisenberg et al. (2000).

A number of genes of unknown from prokaryotic genomes have, in fact, been predicted to interact using the gene neighbor method (Dandekar et al., 1998; Overbeek et al., 1999). It will be of interest to determine if these predicted interactions can be verified experimentally. Nevertheless, the gene neighbor method, as with the other computational methods described above, will be a useful means to guide experimental protein-protein interaction mapping studies.

This page intentionally left blank.

Chapter 7

PROTEIN ARRAYS AND PROTEIN CHIPS

Determining mRNA levels using DNA probes arrayed on chips has become an established method to evaluate gene expression for a variety of tissues and organisms. One format is the Affymetrix oligonucleotide array that consists of oligonucleotides whose sequence is complementary to sequences within each gene of an organism arrayed in high density on a chip. There are multiple oligonucleotides represented for each gene. The second format is a cDNA array consisting of PCR products that include a large portion of each gene arrayed on a glass slide. In this format there is usually one DNA fragment arrayed per gene. For both methods, mRNA expression levels are determined by first isolating RNA from the tissue or organism of interest and attaching a fluorescent label. The labeled RNA is then hybridized to the oligonucleotides or cDNA sequences on the chip. If an RNA is complementary to a sequence on the chip it will be retained and emit a fluorescent signal. Because the precise arrangement of oligonucleotides or cDNAs with respect to gene identity is known, it is possible to quickly determine the pattern of gene expression in a tissue or organism based on the pattern of fluorescent signals from the chip.

A variety of formats for protein arrays are possible. For example, a set of antibodies can be gridded on a filter or slide and used to detect protein expression levels (Pandey and Mann, 2000). Another type of array consists of proteins from an organism arrayed directly on to a glass slide, nylon filter or in microtiter wells (MacBeath and Schreiber, 2000). This format could be used to map protein-protein interactions or to associate a catalytic function with a protein.

The difficulty with protein arrays is that proteins do not behave as uniformly as nucleic acid. Protein function is dependent on a precise, and fragile, three-dimensional structure that may be difficult to maintain in an array format. In addition, the strength and stability of interactions between proteins are not nearly as standardized as nucleic acid hybridization. Each protein-protein interaction is unique and could assume a wide range of affinities. Currently, protein expression mapping is performed almost exclusively by two-dimensional electrophoresis and mass spectrometry. The development of protein arrays, however, could provide another powerful

method to explore protein expression and protein-protein interactions on a genome-wide scale.

7.1 Antibody arrays for protein expression mapping

Overview of an antibody array

One of the goals of proteomics is to examine protein expression levels within and between tissues or organisms. Antibodies have historically been used to detect proteins using methods such as Western blotting and ELISA. It is cumbersome, however, to examine hundreds to thousands of proteins using these methods. An alternative is an antibody array whereby antibodies specific to each protein in the organism being examined are arrayed on a filter or glass slide. Protein expression mapping is then performed by obtaining a crude protein lysate of the tissue or organism of interest and labeling the proteins with a fluorescent tag. The protein mixture is allowed to bind to the antibody array. Those proteins that are expressed in the tissue of interest bind their cognate antibodies on the array. After washing to eliminate non-specific binders, bound protein is detected by the fluorescent tag attached to protein in the original lysate (Fig. 7.1). This method is essentially a high-throughput ELISA experiment. The approach has the advantage that it is not necessary to fractionate the crude protein mixture before binding to the array. In addition, it may be possible automate the procedure to examine multiple samples in parallel. The obvious limitation of this approach is the requirement for antibodies that are specific to each protein in the organism under study.

Obtaining antibodies by immunizing animals with purified proteins is a standard method for obtaining specific antibodies. This approach, however, is prohibitively difficult and expensive for large sets of proteins. An alternative is to isolate specific antibodies using phage display libraries. Phage display of combinatorial antibody libraries have been extensively used to select monoclonal antibodies of a desired specificity without the use of conventional hybridoma technology (Rader and Barbas, 1997). The ability to isolate specific antibodies to a number of different proteins from a single phage display library lends itself to automation and therefore has the potential to provide antibodies for genome scale projects.

Antibody structure and function

The antibody molecule is based on a four-chain structure organized into three functional units (Fig. 7.2) (Padlan, 1993). Two of the units are identical and mediate binding to the antigen; these regions are called the Fab (fragment antigen binding) arms of the antibody. The other unit, Fc

(fragment crystalline), is involved in the effector functions of the molecule. There are two identical heavy chains that span the entire molecule and two identical light chains that are present in the Fab region only (Fig. 7.2). There are five classes of antibody, termed immunoglobulin G (IgG), IgM, IgA, IgD, and IgE (Janeway and Travers, 1994). These classes differ in the Fc region and are associated with different effector functions. The focus for phage display antibody libraries has been on the IgG class (Rader and Barbas, 1997).

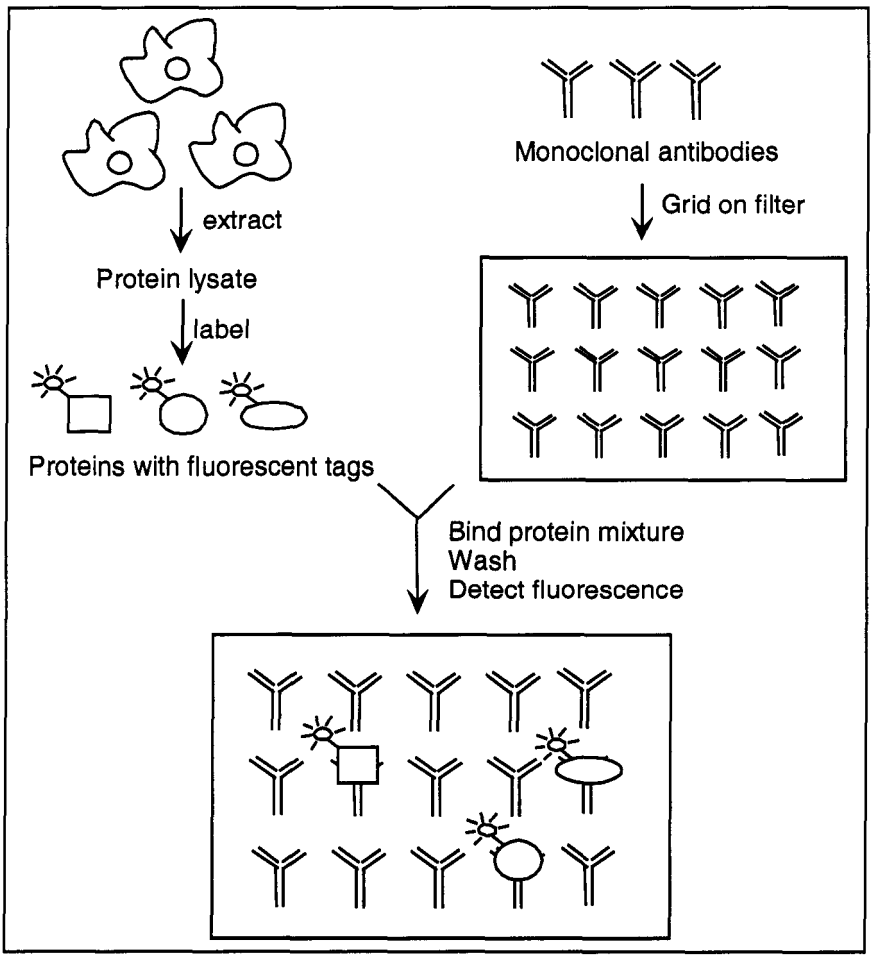


Figure 7.1. Protein expression mapping using an antibody array. The antibody array consists of monoclonal antibodies specific for a set of proteins in the organism of interest gridded onto a filter. To determine if a protein is expressed under the conditions being tested, a crude lysate is obtained and the proteins within the lysate are labeled with a fluorescent tag. The lysate is applied to the filter and the proteins are allowed to bind to the relevant antibody. Bound proteins are visualized via the fluorescent tag.

Cleavage of IgG with papain yields Fab and Fc fragments (Fig. 7.2) (Janeway and Travers, 1994). These fragments are soluble, stable, and retain their molecular functions. The Fv fragment is a sub-fragment of the Fab and consists of just the heavy and light chain variable regions. The Fv fragment is much less stable and is often expressed in recombinant form in *E. coli* with a peptide linker engineered between the V_H and V_L domains to produce single chain Fv (scFv) molecules with improved folding and stability (Barbas et al., 2001). The length of the peptide linker determines the multimerization state of the scFv molecule; long linkers (>20 amino acids) result in monomeric molecules while short linkers (five amino acids) result in dimeric svFv molecules (Hollinger et al., 1993). The multimerization state of the svFv can affect its binding properties in that avidity effects can lead to the dimeric scFv binding to antigens where there is only weak affinity.

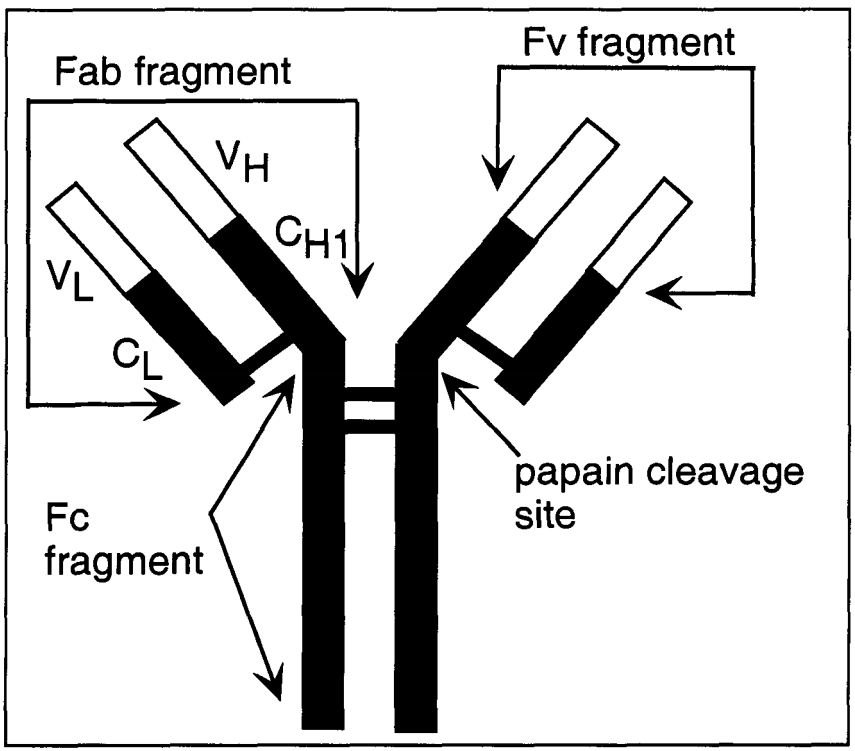


Figure 7.2. Structure of the IgG antibody molecule. The heavy chain and light chain are shown with the disulfide bonds connecting the chains. The Fc fragment is generated by protease cleavage and consists of the C-terminal section of the heavy chains. The Fab fragment consists of the variable and constant regions of the light and heavy chains connected by a disulfide bond. The Fv fragment consists of the variable regions of the light and heavy chains.

Antigen recognition is mediated by a binding site consisting of regions from both the variable heavy (V_H) and variable light (V_L) regions of the Fab (Janeway and Travers, 1994; Padlan, 1993). Within each of the variable chains are non-contiguous hypervariable regions, named complementarity-determining regions (CDRs). The three-dimensional structure of antibodies indicates that the CDRs are loops protruding from a fl-barrel core structure. The CDR loops are juxtaposed in the antibody structure to form the antigen-binding site. The loops are surface exposed and tolerate many different sequences as well as insertions and deletions. This permits the diversity of antigen-binding sites that are used by the immune system to generate antibodies with a vast array of molecular recognition properties.

Phage display antibody libraries

Phage display libraries are constructed by amplifying single-chain Fv (scFv) or Fab antibody fragment genes from the germline using PCR and inserting these genes into phage display vectors as fusions to M13 gene III or gene VIII capsid protein (phage display is described in Chapter 5) (Lerner et al., 1992). A large number of oligonucleotide primer sequences that can be used to amplify Fv or Fab regions have been published for several organisms, including humans and mice (Barbas et al., 2001)(Fig. 7.3). For scFv library construction, the V_L and V_H regions are amplified by an initial PCR reaction and then connected via a linker region by a subsequent overlap extension PCR step (Barbas et al., 2001). The final product encoding the scFv is then inserted into a phage display vector as a fusion to gene III or gene VIII. For Fab library construction, the V_HñC_{H1} and V_L-C_L regions are amplified and inserted into a phage display vector such that the heavy chain region is fused to the capsid gene while the light chain is inserted next to a bacterial signal sequence to direct secretion of the chain to the periplasm of *E. coli.* Within the oxidizing environment of the periplasm, a disulfide bond forms between the heavy and light chains to place the entire Fab fragment on the surface of the phage (Fig. 7.3) (Barbas et al., 2001).

Two types of antibody libraries can be constructed, immune or non-immune. Immune libraries are constructed by immunizing the animal of interest with an antigen(s). In the case of humans, the source can be volunteers with the disease or condition under study (Persson et al., 1991). Human antibodies have also been obtained from severe combined immunodeficiency mice populated with human peripheral blood lymphocytes. Mice are often used because of the extensive use of murine monoclonal antibodies and an extensive range of PCR primers are available for amplifying Fv and Fab fragments. Immunization of rabbits has the advantage that there are fewer antibody genes to amplify by PCR and it is

easier to obtain blood from rabbits. Regardless of the source, mRNA is isolated from peripheral blood and used as a template for PCR amplification of antibody genes.

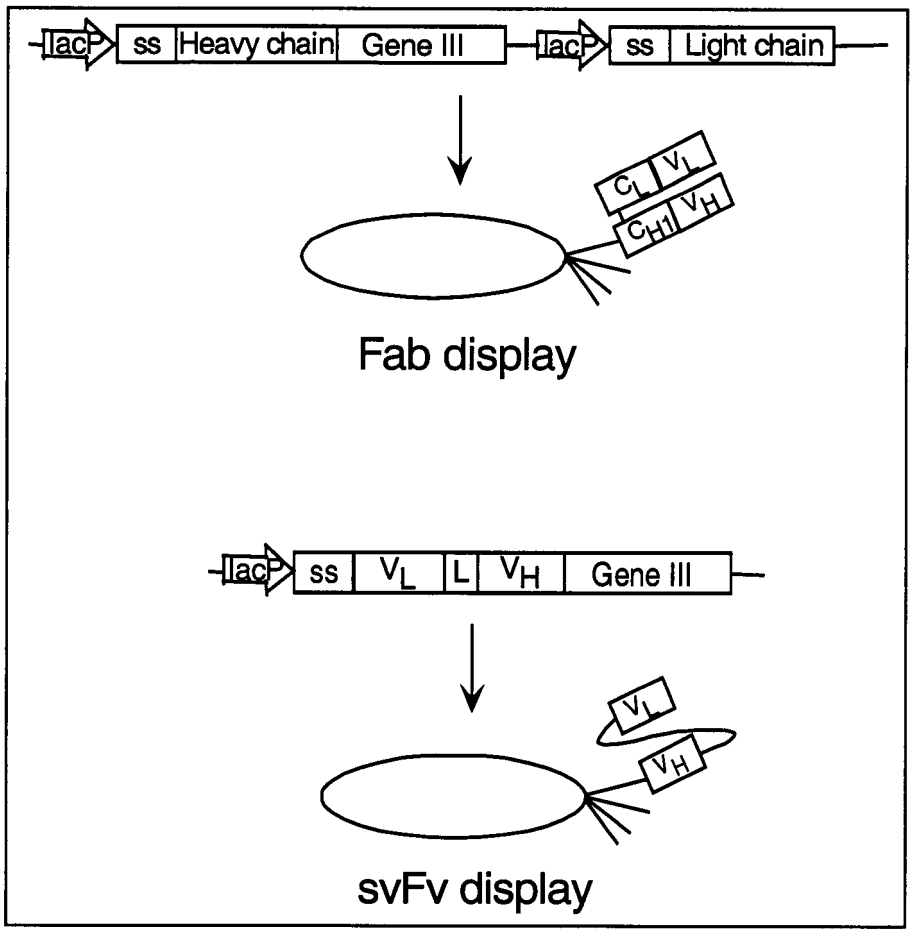


Figure 7.3. Phage display of Fab and Fv antibody libraries. For display of the Fab fragment, the DNA encoding the variable and constant regions of the heavy and light chains are cloned into separate sites of the phagemid vector used for phage display. The chains are fused to a bacterial signal sequence (ss) to direct secretion of the chains to the periplasmic space. The heavy chain is also fused to the M13 bacteriophage gene III protein that mediates attachment to the phage particle. The light chain associates with the phage particle within the periplasmic space via interaction and disulfide bond formation with the heavy chain. The Fv fragment is displayed as a single chain with a peptide linker connecting the V_L and V_H domains.

The advantage of using an immunized animal for PCR amplification, gene cloning and phage display is that there is a high probability of obtaining a specific, high affinity antibody from even modest sized phage libraries. The disadvantage of this approach for proteomics applications is the relatively limited number of protein antigens to which antibodies could be generated. It would be very expensive to immunize animals individually with proteins encoded by all of the open reading frames of an organism. One alternative may be to immunize animals with pools of proteins and isolate specific antibodies by phage display panning in high throughput on the individual proteins.

A more general approach to obtaining monoclonal antibodies to a large set of proteins is to use a non-immune library. In principle, these libraries offer the possibility of obtaining high-affinity antibodies to any protein without the need for immunization (Lerner et al., 1992; Marks et al., 1991). Because the libraries are constructed from a non-immune source, it is essential that the libraries be very large. In general, the affinity of the antibodies isolated from such libraries is proportional to the initial size of the phage library used for selection (Vaughan et al., 1996).

There are several reports of high affinity antibodies obtained from non-immune libraries. For example, an scFv phage display library was constructed by PCR amplification of V-gene segments from 43 non-immunized human donors and insertion of the gene fragments into a phage display vector such as that shown in Fig. 7.3B. A library consisting of 1.4 x 10^{10} scFv fragments was constructed by employing multiple rounds of PCR and cloning of V-genes (Vaughan et al., 1996). The library was used for panning with several different antigens and it was possible to isolate antibodies that exhibit binding affinities of $K_d < 10$ nM (Vaughan et al., 1996). This range of binding affinity is well within the requirements for an antibody array designed to detect protein expression levels, such as that outlined in Fig. 7.1.

Because of the correlation between library size and the ability to select high affinity antibodies, creation of extremely large libraries is an important goal. A combination of cloning of V-gene fragments to create scFv libraries followed by recombination between the V_H and V_L segments within *E. coli* has been used to generate libraries with approximately 100-fold higher diversity than those obtained from cloning alone (Sblattero and Bradbury, 2000). The method involves creation of the initial scFv library with the linker between the V_H and V_L regions also containing a *loxP* sequence. The *loxP* sequence is the recognition site for the Cre recombinase enzyme from P1 phage (Abremski et al., 1983). The initial scFv phage library is introduced at a high multiplicity of infection (MOI) into *E. coli* cells expressing Cre. At the high MOI, multiple scFv clones enter each *E. coli* cell and Cre-mediated recombination occurs between clones at the *loxP*

site to create new combinations of V_H and V_L segments and thus greatly increase the diversity of the library (Fig. 7.4) (Sblattero and Bradbury, 2000). This method has been used to create libraries so large that the complete diversity available in the library cannot be accessed due to limitations in phage production and panning (Sblattero and Bradbury, 2000). The phage libraries have also been used to generate high-affinity antibodies to a number of different proteins (Sblattero and Bradbury, 2000). One of the great advantages of this method is that diversity is generated each time the initial cloned scFv library is introduced into the Cre-producing *E. coli* cells. Therefore, a single cloned primary library could be used virtually indefinitely without losing diversity. In contrast, primary libraries constructed by cloning alone are a limited resource, as amplification cannot be carried out without a potential loss in diversity.

Despite the availability of large, non-immune phage display libraries, it is a challenge to screen the library for antibodies to hundreds of different proteins. Improvements in screening technology, however, are rapidly removing this bottleneck (de Wildt et al., 2000). For example, a high-throughput screen for recombinant antibodies via a protein array has recently been reported (de Wildt et al., 2000). The method involves robotic picking and high-density gridding of bacteria containing antibody genes followed by filter-based ELISA to identify clones that bind the antigen of interest. For these experiments, a phage antibody library was used for one round of panning on the antigen of interest. Phage that bound to the antigen were eluted and used to infect *E. coli.* A robot was then used to pick colonies for gridding at high density. The method was used to screen up to 18,342 clones simultaneously for antibodies that specifically bind an antigen (de Wildt et al., 2000). Importantly, the antibody screen can be performed in parallel for a number of antigens. This work also clearly demonstrates the feasibility of arraying antibodies at high density for protein expression mapping experiments.

In summary, large non-immune antibody libraries coupled with high-throughput screening techniques should permit the isolation of recombinant antibodies on a genome-wide scale. These antibodies will serve as powerful tools for protein expression mapping using antibody arrays (Fig. 7.1). In addition, these antibodies can be used to identify spots from 2-D gels and thus facilitate protein expression mapping by 2D gel electrophoresis.

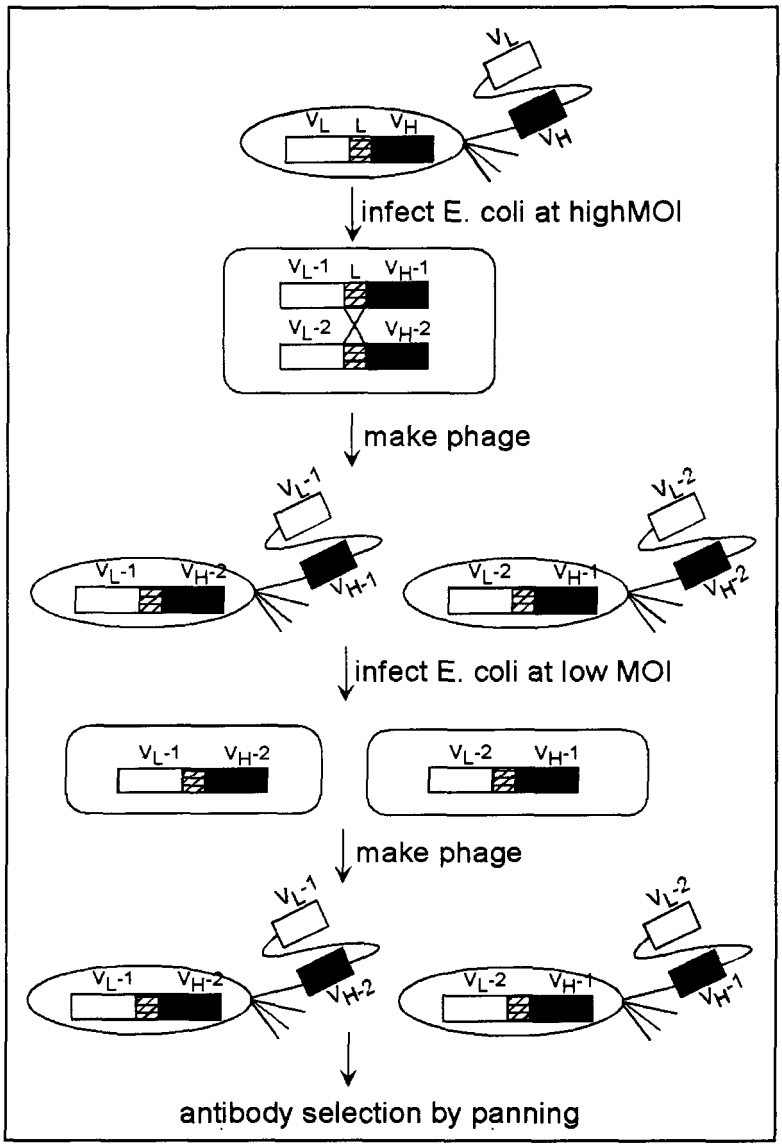


Figure 7.4. Creation of a complex antibody library by shuffling of V_L and V_H genes in an scFv library using Cre-lox recombination. The peptide linker connecting the V_L and V_H domains also encodes a cis-acting recombination sequence for the Cre recombinase. The library is constructed by cloning V_L and V_H genes into the phage vector and producing a phage library. The library is used to infect *E. coli* at high multiplicity, which results in multiple clones within each cell. The *E. coli* cells express the Cre recombinase that catalyzes shuffling of the V_L and V_H genes. A phage library is again produced and the phages are used to infect *E. coli* at a low multiplicity so that only one clone is present in each cell. The final phage library is then produced for panning on the antigen of interest.

7.2 Functional analysis using peptide and protein arrays

Overview

One of the goals of protein chip technologies is to array all of the proteins encoded within a genome for functional studies. As described in Chapter 4, systematic efforts are underway to construct defined sets of cloned genes for high-throughput expression and purification of recombinant proteins (Hartley et al., 2000; Hudson et al., 1997). As these efforts mature, it will be feasible to construct protein arrays for systematic and global analyses that provide information on the biological function of unclassified proteins.

Global analysis of protein function has historically been studied using library screens. For library screens, sets of related elements, such as a cDNA expression library, are tested in batch for a desired biological property. For example, screening of expressed proteins in A-phage plaques has been performed for many years (Young and Davis, 1983). For these experiments, a cDNA library is inserted into a A-phage vector such that the encoded protein is expressed as a fusion protein with the *E. coli* β-galactosidase protein. The phage library is used to infect *E. coli* and A-plaques are formed on agar plates. Each A-plaque expresses a different fusion protein. The plaques are then probed for the presence of specific proteins using antibodies specific for those proteins (Young and Davis, 1983). When a positive plaque is identified, it is picked and the DNA sequence of the insert is determined to identify the relevant gene. In this way, it is possible to clone genes based on a screen for an antibody-antigen interaction.

Protein arrays may offer advantages over libraries. The array format provides a precise spatially oriented grid that allows side-by-side comparison of assay results for all of the proteins on the array. The spatial arrangement also permits immediate identification of a clone that tests positive in an assay based on its location on the array. Therefore, less effort is required to identify the protein responsible for an interaction than with a library screen. A disadvantage of a protein array, however, is that fewer proteins can be efficiently arrayed and screened (~10^4) versus a library (~10^9). Therefore, the number of elements that can be effectively arrayed and assayed currently limits protein arrays.

Peptide arrays

The earliest applications of the array format have been for peptide arrays. For example, the pin method for peptide synthesis involves the

parallel synthesis of peptides in a 96 well microtiter plate format. Peptides are synthesized on 4 mm pins using a fluorenylmethoxycarbonyl (Fmoc) amino acid protection strategy (Geysen et al., 1984; Geysen et al., 1986). For each cycle of synthesis, the pins are immersed in an appropriate amino acid solution. Because the peptides are synthesized on a solid support, they can be washed extensively between cycles of addition to increase the purity of the final preparation. This method was initially used for epitope-mapping on the VP1 coat protein of foot-and-mouth disease virus (Geysen et al., 1984; Geysen et al., 1986). All of the 208 possible overlapping hexapeptides present in the 218 amino acid VP1 sequence were synthesized and assayed for interaction with an antibody. In this way an immunodominant region was identified and subsequently fine-mapped by introducing single amino acid substitutions into a peptide encompassing the region (Geysen et al., 1984). This was one of the first uses of high-throughput methods to define protein-protein interactions.

The SPOT-synthesis method also employs Fmoc chemistry but uses hydroxyl groups present on cellulose filter paper to derivatize and thereby immobilize β-alanine groups onto the paper. After deprotection, the β-alanine groups can be used as platforms for the synthesis of peptide arrays (Fig. 7.5) (Frank, 1992; Gausepohl et al., 1992). This method has been widely used for mapping antigen-antibody interactions as well as protein-DNA, protein-metal and other protein-protein interactions (Reineke et al., 2001).

The most frequent application of SPOT-synthesis has been in the preparation of peptide arrays for the identification of linear B-cell epitopes. If the protein antigen is known, a set of overlapping peptides that encompass the entire sequence can be readily synthesized and assayed for binding of antibody (Reineke et al., 1999). The individual residues critical for binding can then be determined by SPOT-synthesis of peptides containing amino acid substitutions.

Mapping of discontinuous epitopes is more difficult because of the low affinity for antibody binding of peptides derived from separate regions of the antigen protein. An interesting application of SPOT-synthesis has been to map a discontinuous binding site on interleukin-10 (IL-10) for a neutralizing anti-IL-10 antibody called CB/RS/1 (Reineke et al., 1999). An overlapping peptide scan of the IL-10 sequence was performed using 15-mer peptides shifted by one amino acid. The peptide array was then probed using the CB/RS/1 antibody and bound antibody was detected using an anti-mouse IgG peroxidase-labeled antibody. The CB/RS/1 antibody was found to bind to peptides representing two regions of the protein that are distant in the primary sequence but continuous on the folded structure (Reineke et al., 1999). Amino acid residue positions within peptides that did bind the antibody were then systematically substituted and arrayed by SPOT-

synthesis. Binding of the CB/RS/1 antibody was again assayed to determine the positions within the peptides that are critical for antibody binding. A number of substitutions were also found that appeared to increase antibody binding. The two regions that exhibited antibody binding were linked into a single peptide that also incorporated the substitutions that increased binding. The single peptide was then systematically substituted with cysteine residues to identify disulfide bonds that would increase binding by lowering the conformational entropy of the peptide. The end result of these multiple SPOT-synthesis experiments was a tight-binding 32-mer-peptide mimic of the discontinuous binding site between IL-10 and the CB/RS/1 antibody (Reineke et al., 1999). This study clearly demonstrates the power of the spot synthesis method for the rapid construction and testing of peptide arrays.

A number of other protein-protein interactions have been studied using SPOT-synthesis. These include a number of interactions between proteins involved in signal transduction such as PDZ domains (Schultz et al., 1998), SH3 domains (Cestra et al., 1999) and tumor necrosis factor receptor-associated factors (TRAFs) (Pullen et al., 1999). In addition, the method has been used to define the substrate specificity of the bacterial chaperone protein, SecB (Knoblauch et al., 1999). This study is of particular interest because chaperone proteins exhibit relaxed substrate specificity and thereby bind to a large number of proteins. For these experiments, a total of 2688 peptides derived from 23 different proteins were synthesized and screened for SecB binding (Knoblauch et al., 1999). The results were used to identify a recognition motif that can be used to accurately predict SecB binding peptides. These experiments represent an important step towards the use of SPOT-synthesis for proteome-wide screens of binding specificity. SPOT-arrays containing overlapping peptides for an entire bacterial proteome may soon be generated. Such an array would be a powerful tool for the study of antigen-antibody and protein-protein interactions at the level of the proteome.

Although the SPOT-synthesis method has great utility, the density of spots on the filters is not of the same order as arrays on DNA chips or glass slides (Brown and Botstein, 1999). It has been shown, however, that photolithography can be used to greatly increase the density of peptides on an array (Fodor, 1991). This method is similar to that used for the construction of highdensity oligonucleotide arrays. Photolabile protecting groups are used for peptide synthesis. Growing peptides are selectively deprotected using masks that allow light to reach only those peptides to which an amino acid is to be added. Peptide densities up to 250,000 per cm^2 have been achieved (Fodor, 1991).

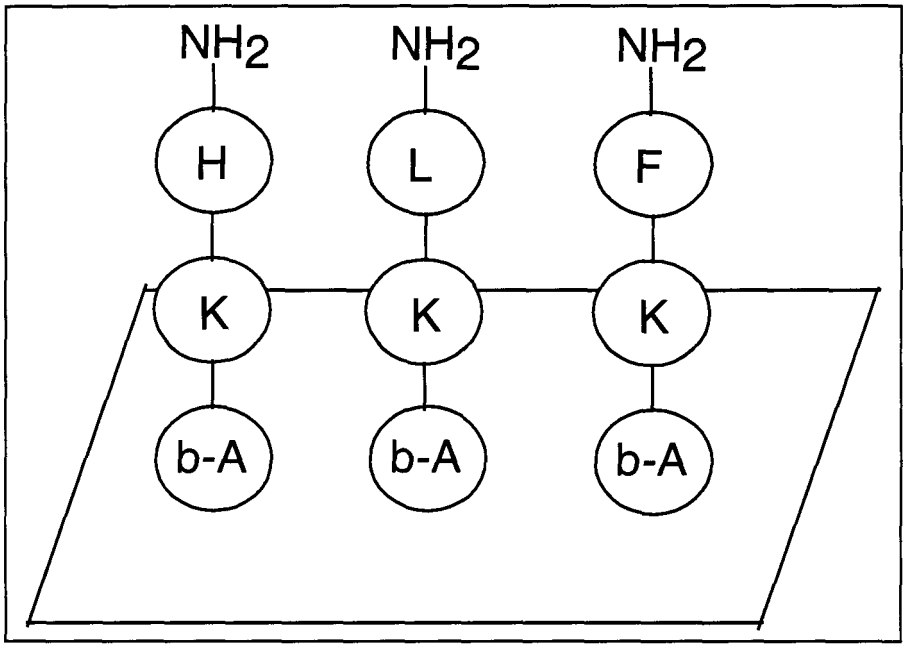


Figure 7.5. Peptide array construction by SPOT-synthesis. fl-alanine groups (b-A) interact with the cellulose filter that serves as a planar support. Peptide synthesis then proceeds using Fmoc chemistries using the fl-alanine group as a starting point. The peptide is attached to the filter via its carboxy-terminus. In this case, lysine is added at the second position and various amino acids are present at the amino terminus of the peptide.

The arrays described above are constructed by the synthesis of peptides directly onto a solid support. At present, SPOT-synthesis is limited to approximately 30-40 amino acids due to difficulties with product purities and coupling efficiencies (Molina et al., 1996). Therefore, construction of protein arrays by direct synthesis is not possible. Instead, proteins must be expressed and purified subsequent to use in an array. Recombinant proteins can often be expressed in organisms such as *E. coli, S. cerevisiae,* or from tissue culture. Alternatively, proteins can be produced by *in vitro* transcription and translation. The difficulty with any of these methods is obtaining properly folded protein for use on the array. Improper folding and aggregation is a common problem for recombinant protein expression in heterologous systems. In addition, *in vitro* expression of proteins in the absence of specific chaperone proteins may lead to improperly folded structures. This problem is the major factor that distinguishes protein arrays from DNA or oligonucleotide arrays. Thus, progress in the development of protein arrays has been slower than that for DNA arrays. Nevertheless,

several recent developments suggest protein arrays may be possible.

Biochemical genomics: Assay of pooled recombinant proteins

The first obstacle in the development of a protein array is the large-scale expression and purification of proteins encoded by the open reading frames (ORFs) of an organism. The feasibility of this approach has been demonstrated for *S. cerevisiae* (Martzen et al., 1999). For these experiments, an array of 6144 yeast strains was constructed. Each strain contained a plasmid expressing a different glutathione-S-transferase (GST)-ORF fusion under the transcriptional control of the P_{CUP1} promoter. Because it would be prohibitively difficult to purify 6144 individual proteins, the strains were collected into 64 pools consisting of 96 different GST fusion strains in each pool (Fig. 7.6). The GST fusion proteins from each of the pools were purified in batch by affinity chromatography with a glutathione agarose resin (Martzen et al., 1999). The pools were then used for biochemical assays of protein function. For example, assay of the GST fusion pools demonstrated that each of the two previously known tRNA splicing activities from yeast were present only in those pools expected to contain the genes encoding these activities based on ORF number.

When a pool was identified as containing a biochemical function, the individual strain responsible for the activity was determined by preparing and assaying the GST fusion proteins from each of the 8 rows and 12 columns of strains from the appropriate microtiter plate (Martzen et al., 1999). Using this approach, biochemical assays assigned a putative function to the proteins encoded by three previously unknown genes. These included two cyclic phosphodiesterases involved in tRNA processing and a methyltransferase capable of modifying cytochrome c (Martzen et al., 1999).

In principle, the use of addressable, pooled GST fusion proteins could be used to identify proteins associated with any biochemical activity, assuming that the fusion protein is soluble, folded and functional. The method has the additional advantage that, once the GST fusion clones are constructed, it is a rapid technique. The authors state that only two weeks are required to purify the 64 pools and the assays can be accomplished in a day (Martzen et al., 1999). In addition, the method is sensitive because only 96 recombinant proteins are assayed at one time in contrast to the use of cell lysates where thousands of proteins are present. This leads to a much higher concentration of each protein, which greatly facilitates detection of a biochemical activity (Martzen et al., 1999).

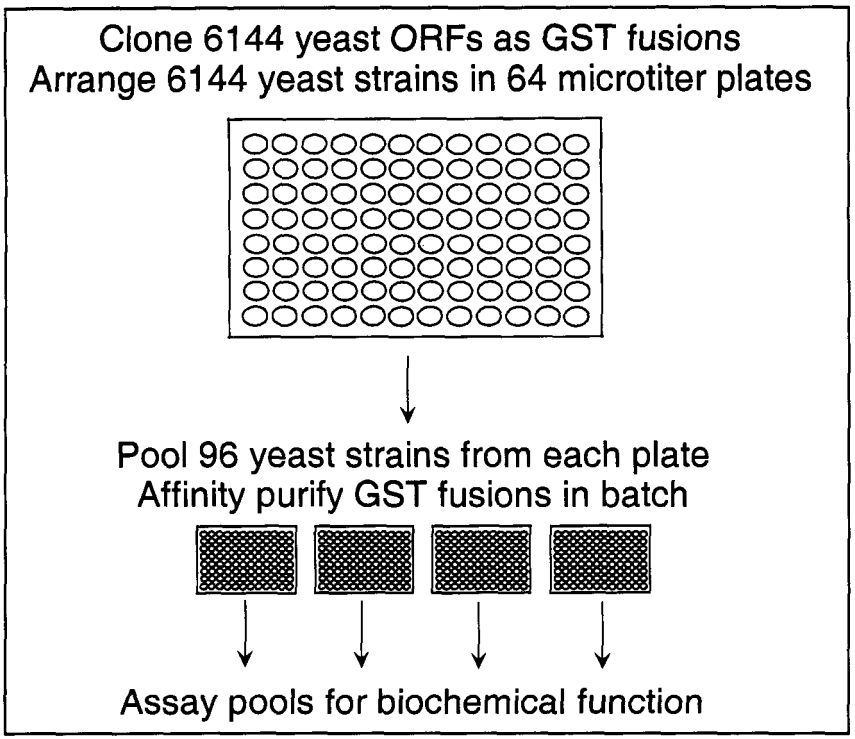


Figure 7.6. Purification of protein from pooled yeast strains. Each yeast ORF was cloned as a fusion to glutathione-S-transferase in a protein expression vector to create 6144 yeast strains. The individual strains were pooled in groups of 96 to create a set of 64 pools. Each pool was grown and the 96 fusion proteins are purified in batch. Each pool was then assayed for a biochemical function (Martzen et al., 1999). Pools positive for function were then deconvoluted using smaller pools consisting of strains from rows and columns of a 96-well plate.

The power of the pooled GST fusion protein approach will increase as new biochemical reagents and assays become available. The development of chemical probes for biological processes, termed chemical biology, is a rapidly advancing field. For example, the chemical synthesis of an active site directed probe for identification of members of the serine hydrolase enzyme family has recently been described (Liu et al., 1999). The activity of the probe is based on the potent and irreversible inhibition of serine hydrolases by fluorophosphate (FP) derivatives such as diisopropyl fluorophosphate. The probe consists of a biotinylated long-chain fluorophosphonate, called FP-biotin (Liu et al., 1999). The FP-biotin was tested on crude tissue extracts from various organs of the rat. These experiments showed that the reagent can react with numerous serine hydrolases in crude extracts and can detect enzymes at subnanomolar

concentrations (Liu et al., 1999). Clearly, reagents such as FP-biotin would work well with the pooled GST fusions where the proteins are present at a higher concentration than in crude extracts (Martzen et al., 1999). Other such chemical probes will likely be developed in the next few years.

Protein microarrays

The use of pooled GST fusion proteins allows tests of biochemical activity but does not lend itself to the identification of protein-protein or protein-small molecule interactions. For these purposes, it is necessary to immobilize proteins on a solid support so that non-binding molecules can be washed away. It is also necessary that the protein, once attached to the solid support, retain its folded conformation. Several reports of protein microarrays have been published. For example, recombinant proteins from cDNA expression libraries have been spotted at high density on polyvinylidene difluoride filters (PVDF) (Lueking et al., 1999) (Fig. 7.7). The density was such that a filter the size of a microscope slide (25 x 75 mm) contained 4800 samples. By spotting known amounts of purified human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and detecting the protein with a monoclonal anti-GAPDH antibody, it was shown 250 attomol (10^{-18} mol) of a spotted protein could be detected with the method (Lueking et al., 1999). Protein expression from the spotted cDNA library was detected using an antibody to a poly-histidine-tag present on all of the clones. These experiments demonstrated reliable detection of protein expression with a low rate (11%) of false positives (Lueking et al., 1999). Although specific antibodies were used for detection of proteins, the authors point out that the filters could be screened for interactions with a number of other molecules including other proteins, nucleic acids or small molecules (Lueking et al., 1999).

Another approach has been to immobilize proteins within arrays of microfabricated polyacrylamide gel pads (Arenkov et al., 2000). Nanoliters of protein solutions are transferred to 100 x 100 x 20-μM gel pads and assayed with antibodies that are labeled with a fluorescent tag. Antigen imbedded in the gel pads can be detected with high sensitivity and specificity (Arenkov et al., 2000). Furthermore, enzymes such as alkaline phosphatase can be immobilized in the gel pads and enzymatic activity is readily detected upon the addition of an indicator substrate. The main advantage of the use of the threedimensional gel pad for fixation of proteins is the large capacity for immobilized molecules. In addition, the pads in the array are separated from one another by a hydrophobic surface. Thus, each pad behaves as a small test tube for assay of protein-protein interactions and enzymatic reactions (Arenkov et al., 2000). The disadvantage of the method is the need to microfabricate the array of gel pads in that microfabrication is

not a widely accessible technology.

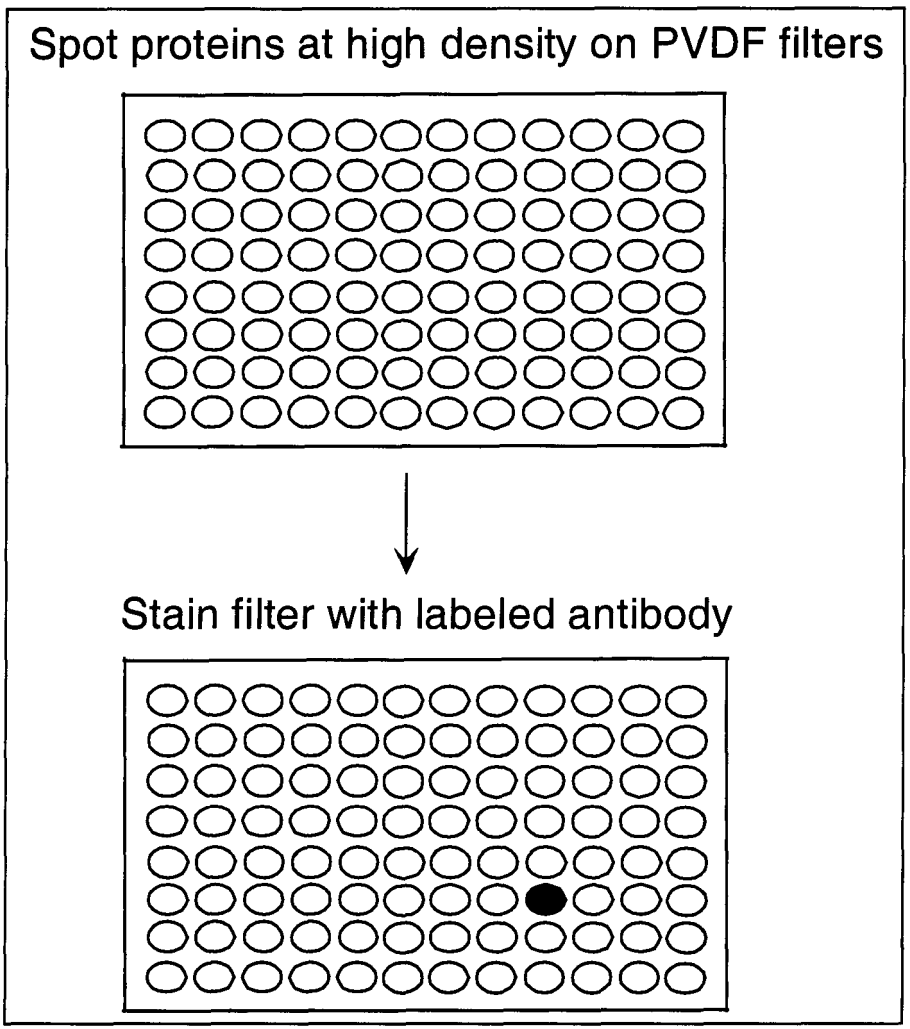


Figure 7.7. Protein microarray on polyvinylidene difluoride filters (PVDF). Proteins are gridded on filter and probed with an antibody to a gene of interest. Protein-ligand interactions could be detected using a labeled ligand as a probe.

The potential of photolithography for the construction of protein microarrays has also been demonstrated (Mooney et al., 1996). For these experiments, antibodies were assembled in precise two-dimensional patterns on silicon wafers. This was accomplished by first forming a self-assembled monolayer of n-octadecyltrimethoxysilane (OTMS) on a silicon-dioxide

surface. Sections of the monolayer were removed by W photolithography to generate a pattern on the surface. The OTMS coated regions were shown to adsorb bovine serum albumin (BSA) while the uncoated regions did not (Mooney et al., 1996). Other proteins could be coated onto the OTMS regions by adsorbing BSA that had been conjugated with biotin and then adding streptavidin to bind the biotin-BSA. The immobilized streptavidin was then used to bind a biotinylated protein of interest (Mooney et al., 1996). Note that this is possible because each streptavidin molecule contains four biotin-binding sites. This method has been clearly demonstrated for the immobilization of a single protein in a pattern on a chip but it has not yet been used to construct a microarray consisting of many different proteins.

A method has recently been described for the attachment of proteins to glass slides at high spatial densities (MacBeath and Schreiber, 2000). These protein microarrays were constructed using a high-precision robot to deliver nanoliter volumes of samples to glass slides at a density of 1600 spots per square centimeter. The proteins were attached covalently by pre-treating the slides with an aldehyde containing silane reagent (MacBeath and Schreiber, 2000). The aldehydes react readily with primary amines from lysine residues as well the α-amine at the NH_2-termini of the protein. Because lysines are usually present at multiple positions on the surface of proteins, the molecules attach in multiple orientations.

The highdensity protein microarray was used to examine protein-protein interactions using three pairs of proteins that are known to interact: protein G and IgG; p50 and IkBa; and the FKBP12-rapamycin binding domain (FRB) and FKBP12 (MacBeath and Schreiber, 2000). For each of these experiments, one of the binding partners was immobilized while the other was labeled with a fluorescent tag and allowed to bind to the slide. Bound protein was detected by retention of the fluorescent tag on the slide after extensive washing (MacBeath and Schreiber, 2000). These experiments demonstrated that the microarray can be used to efficiently detect protein-protein interactions from extremely small sample volumes. The immobilized proteins were spotted at a concentration of 100 μg/ml. Because the binding reaction takes place in nanoliter volumes, the amount of the solution phase protein needed to detect binding is very small. For instance, specific binding could be detected for the FRB-FKBP12 interaction using picogram quantities of FXBP12 (MacBeath and Schreiber, 2000). The authors note that, because the concentration of solution phase protein necessary for binding is so low, it may be possible to label proteins with a fluorescent tag directly in a cell lysate and use the array to quantitate the amount of a specific protein within the lysate. Thus, the protein microarray could be used for protein expression mapping as well as for detecting protein-protein interactions (MacBeath and Schreiber, 2000). Using a similar approach, it was shown that the binding of small molecules to proteins on the microarray could be efficiently assayed. At this point it is unclear what fraction of

proteins will retain a folded structure when immobilized but based on initial experiments, the protein microarray has great potential for high throughput protein assays.

Protein chips and mass spectrometry

The use of mass spectrometry for the identification of proteins was described in Chapter 2. Because of the speed and accuracy of mass spectrometry, there is a strong interest in coupling the advantages of the highly parallel aspects of protein arrays with efficient identification of proteins by mass spectrometry. For example, the use of MALDI-TOF mass spectrometry in combination with a protein array has been described for the analysis of amyloid β peptide variants secreted from tissue culture cells (Davies et al., 1999). Aggregated forms of the 4-kDa amyloid β peptide form the senile plaques that are often found in the brain tissue of patients suffering from Alzheimer's disease (Selkoe, 1998). Numerous variants of the peptide have been identified in clinical samples and therefore an efficient method for identifying the variants is required. For this purpose, a protein chip was constructed whereby an antibody to the amyloid β peptide was immobilized on the chip surface (Davies et al., 1999). Amyloid β peptide variants secreted from cultured cells were captured from the media by placing 1 μl of media onto the chip surface. The chip was then washed and bound peptide was eluted and analyzed by MALDI-TOF mass spectrometry (Davies et al., 1999). The high sensitivity and accuracy of mass spectrometry allowed the accurate identification of several amyloid β peptide variants. In addition, a controlbovine IgG was immobilized at a different position on the chip to show that the antibody to the amyloid peptide was responsible for capturing the variants from the media (Davies et al., 1999). This study demonstrates that the combination of a protein chip with mass spectrometry is an efficient means of identifying peptides with subtle differences in composition.

The protein chip-mass spectrometry approach has been expanded to include a number of immobilization platforms for molecules of interest. Thus, molecules can be captured by an affinity method such as an antibody, or using chip surfaces with other chromatographic properties such as anion exchange, cation exchange, metal affinity or reverse phase (Fung et al., 2001). These chips can be used to reduce a complex mixture of proteins to sets of proteins with common properties that are then analyzed by mass spectrometry. In concept, this approach is similar to the liquid chromatography (LC) and tandem mass spectrometry approach described in Chapter 2 (Link et al., 1999). The goal of both approaches is to reduce the complexity of proteins to a number that can accurately be examined by mass spectrometry. The spectrum of proteins identified using a given

chromatographic fractionation method represents a fingerprint of the cell state when the protein was extracted. For example, the protein chip method has been used to study protein expression profiles in normal and cancerous prostate samples to identify protein markers characteristic of a disease state (Wright et al., 2000). The protein chip-mass spectrometry platform is commercially available and a number of other applications have been published (reviewed in (Fung et al., 2001)).

Use of DNA microarrays to study proteinfunction

DNA microarrays are widely used to study gene expression by measuring mRNA levels (Brown and Botstein, 1999). An interesting alternative is to use these arrays to study the DNA-binding specificities of transcription factor proteins. A DNA microarray-based method has been developed to characterize sequence-specific DNA recognition by zinc-finger proteins (Bulyk et al., 2001). The zinc-finger transcription factors are among the best-understood families in terms of sequence-specific DNA binding. The Zif268 transcription factor from the mouse was used as a model system for these studies. The experiment involved placing the Zif268 protein on the surface of the M13 filamentous phage to isolate a number of variants with altered binding specificity by mutagenesis and phage display screening. The binding specificity of the wild type and mutant Zif268 proteins was then determined using a DNA microarray.

The Zif268 protein contains three zinc fingers (F1, F2, F3) and each of these fingers interacts with a three base pair sequence (Pavletich and Pabo, 1991). Amino acids within the F2 finger were mutagenized and screened by phage display to isolate binding variants. A DNA microarray was constructed that contained all of the 64 possible combinations of the 3 base pair binding region of the F2 zinc finger in addition to the flanking wild type F1 and F3 recognition sequences (Bulyk et al., 2001) (Fig. 7.8). Phage displaying wild type or mutant Zif268 protein were allowed to bind the DNA sequences on the array, non-binders were washed away, and bound phage were detected with an anti-M13 antibody labeled with a fluorescent tag (Bulyk et al., 2001). Because the binding assay is highly parallel, it was possible to obtain a complete description of the binding specificity of each mutant in a single microarray experiment.

It should be possible to extend the DNA microarray-binding experiment to whole-genome analysis of transcription factor binding sites. The authors suggest that a microarray spotted with 12,000 one-kilobase sequences would span the entire *Saccharomyces cerevisiae* genome (Bulyk et al., 2001). Such an array could be used to characterize the sequence specificity of *S. cerevisiae* transcription factors. These experiments would be useful for predicting functions of previously uncharacterized transcription

factors as well as identifying new regulatory networks (Bulyk et al., 2001).

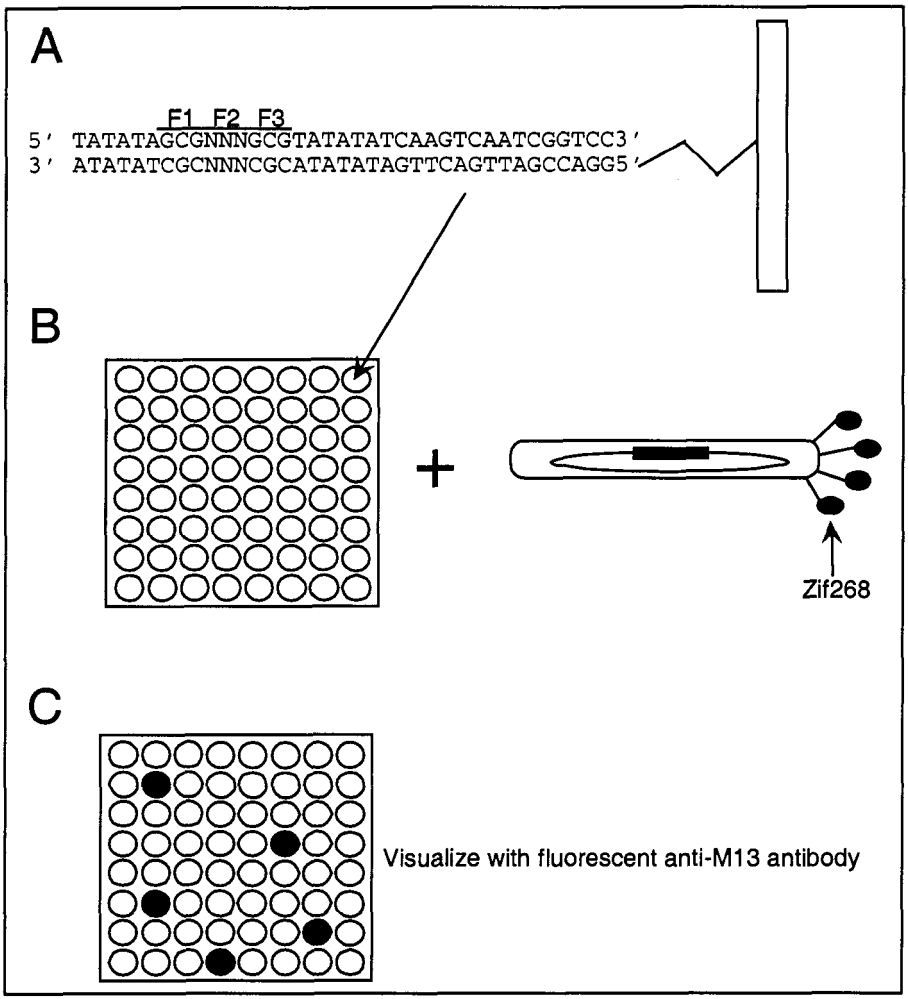


Figure 7.8. Use of DNA microarray to define the binding specificity of transcription factors. **A.** Immobilized DNA fragment containing the Zif268 binding site. The binding site for the F2 finger is represented as NNN where N indicates any nucleotide. **B.** A microarray containing the sequence in (A) with all 64 combinations of the three base pair F2 binding site was constructed and phages displaying a variant Zif268 protein were allowed to bind. **C.** Bound phages were detected with an anti-M13 antibody. The position of bound phage defines the substrate specificity of the Zif268 protein variant (Bulyk et al., 2001).

6.3 Surface plasmon resonance biosensor analysis

Overview

Surface plasmon resonance (SPR) biosensors have become an established method to measure molecular interactions. SPR biosensor experiments involve immobilizing one reactant on a surface and monitoring its interaction with another molecule in solution. SPR is an optical phenomenon used to measure the change in refractive index of the solvent near the surface that occurs during complex formation or dissociation (Jonsson et al., 1991) (Fig. 7.9). The SPR signal is expressed as resonance units (RU) and is proportional to the mass of the molecule in solution interacting with the immobilized ligand. Therefore, the standard experiment involves immobilizing the low molecular weight ligand and detecting the change in signal that occurs when the partner molecule is passed over the immobilization surface in a continuous flow (Schuck, 1997). One of the advantages of SPR is that binding reactions are monitored in real time without the need to label ligands. Hence, SPR can be used to study interactions between proteins, carbohydrates, nucleic acids, lipids and small molecules.

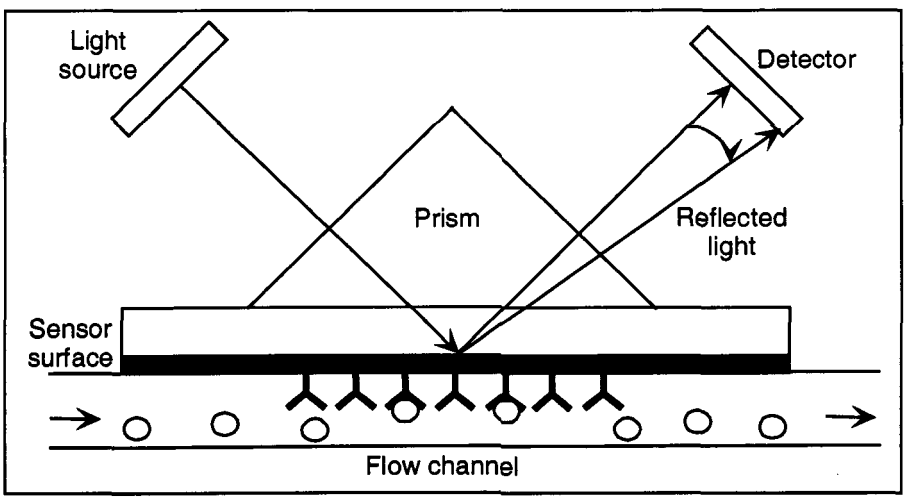


Figure 7.9. Schematic diagram of a surface plasmon resonance biosensor. One of the binding partners is immobilized on the sensor surface. With the BIACORE instrument, the soluble molecule is allowed to flow over the immobilized molecule. Binding of the soluble molecule results in a change in the refractive index of the solvent near the surface of the sensor chip. The magnitude of the shift in refractive index is related quantitatively to the amount of the soluble molecule that is bound.

Measuring interactions of biomolecules by SPR

Several hundred studies on macromolecular interactions using SPR biosensors have been published in a variety of fields (Rich and Myszka, 2000). Many of these studies are focused on detecting and quantitating protein-protein interactions. A typical experiment consists of covalently attaching one of the proteins to the sensor surface. A number of surfaces are commercially available for attachment including carboxymethyl dextran, which can be derivatized to give a number of different functional groups to allow for a variety of immobilization chemistries (Schuck, 1997). Other surfaces include streptavidin for capture of biotinylated molecules and a nickel chelation surface for capture of poly-histidine-tagged proteins (Rich and Myszka, 2000). Binding of soluble protein to the immobilized protein results in an SPR signal in real time that can be used to monitor association kinetics. A buffer solution lacking protein is then allowed to flow over the complex to monitor dissociation kinetics. The kinetic data (k_a, k_d) can be used to obtain an equilibrium constant (K_D) (Morton and Myszka, 1998). Recent developments in SPR instrumentation may permit high throughput analysis of protein-ligand interactions. For example, BIACORE has developed an instrument that can analyze samples from a 96-well plate format. Thus, SPR is likely to make an important contribution to genome-scale protein-protein interaction mapping.

New developments in immobilization surfaces have lead to the use of SPR biosensors to monitor protein interactions with lipid surfaces and membrane-associated proteins. Commercially available (BIACORE) hydrophobic and lipophilic sensor surfaces have been designed to create stable membrane surfaces. It has been shown that the hydrophobic sensor surface can be used to form a lipid monolayer (Evans and MacKenzie, 1999). This monolayer surface can be used to monitor protein-lipid interactions. For example, a biosensor was used to examine binding of Src homology 2 domain to phosphoinositides within phospholipid bilayers (Surdo et al., 1999). In addition, a lipophilic sensor surface can be used to capture liposomes and form a lipid bilayer resembling a biological membrane.

An interesting application of the lipid monolayer on the hydrophobic sensor surface has been the attachment of major histocompatibility complex (MHC) molecules to the surface in a specific orientation (Celia et al., 1999). This was accomplished by incorporating into liposomes a chemically modified lipid containing a nickel salt at the polar head group position (Celia et al., 1999). The liposomes containing the nickel lipids were used to coat the hydrophobic sensor surface to form a monolayer. The recombinant MHC molecule used for the experiments contained a poly-histidine tag in place of the transmembrane spanning region. It was demonstrated by electron microscopy that the histidine-tagged MHC molecules bind to the surface of

the liposome. SPR measurements then demonstrated that the interaction with the liposome observed by EM is due to the histidine-tagged MHC molecule binding specifically to the surface of the lipid bilayer (Celia et al., 1999). Furthermore, SPR experiments demonstrated that the MHC protein interacted with the monolayer in a specific orientation. This was inferred from the observed binding of an anti-$\alpha 1\alpha 2$ antibody to the MHC molecules but the failure to detect binding of an anti-histidine-tag antibody (Celia et al., 1999). The failure of the anti-histidine-tag antibody to bind the MHC molecules was interpreted as steric hindrance due to the histidine-tag being associated with the nickel lipid at the monolayer surface. Finally, by incorporating a fluorescently labeled lipid into monolayers, the authors were able to show that the lipids within the monolayers are laterally mobile (Celia et al., 1999). Thus, sophisticated SPR applications are being developed to simulate membrane protein interactions *in vitro.* Further advances in reconstituting membranes and membrane proteins on sensor surfaces will be an exciting contribution to proteomics studies. Nearly one-third of the open reading frames in a genome are thought to encode membrane-associated proteins. Therefore, the ability to assay and quantitate interactions of proteins within membranes will be crucial to proteomics studies whose goal is to obtain genome-wide protein-protein interaction maps.

lntegration of SPR biosensors with mass spectrometry

As described above, SPR biosensors can be used to detect and quantify protein-protein interactions without the need to label either of the binding partners. The method has also been used to detect interacting proteins from complex mixtures including cell lysates and conditioned media. This approach, known as "ligand fishing", involves immobilizing a known protein as a functional hook to fish unknown binding partners from complex mixtures (Lackman et al., 1996; Nelson et al., 2000). The difficulty with this approach is that, once a potential binding partner is identified in a mixture, the protein must be purified from the mixture using standard biochemical techniques before it can be identified by amino acid sequencing (Lackman et al., 1996). Coupling the biosensor with mass spectrometry provides a more direct route to identification of a binding partner. For this approach, the biosensor serves as micropurification platform for protein identification by mass spectrometry analysis (Williams and Addona, 2000). The amount of protein recoverable from a SPR sensor surface is low (femtomoles), but this is a sufficient amount for identification using sensitive MALDI-TOF or tandem mass spectrometry methods (Nelson et al., 2000; Williams and Addona, 2000).

The potential of coupling SPR with mass spectrometry has been demonstrated using glutathione-S-transferase (GST) and an anti-GST

antibody as a model system (Nelson et al., 1997). For these experiments, anti-GST antibody was coupled to a carboxymethyl dextran chip and free GST was injected and allowed to bind the antibody. The analysis was then stopped; the chip was removed from the machine and the area of the chip containing bound protein was coated with matrix material for MALDI analysis (Nelson et al., 1997). The chip was then analyzed by MALDI-TOF mass spectrometry and the resulting spectrum revealed a mass consistent with the mass of the GST protein. This method has been extended to include proteolytic digestion of samples by using multiple flow cells on a single chip. For these experiments, anti-human interleukin alpha (anti-IL-1α) antibody was immobilized in flow cell-1 of the sensor chip while the protease pepsin was immobilized in flow cell-2 (Nelson et al., 2000). A solution containing IL-1α was routed to flow cell-1 where SPR measurements indicated it was bound by the anti-IL-1α antibody. Following washing of non-specifically bound proteins, the IL-1α was eluted from the surface and routed from flow cell-1 to flow cell-2. After allowing sufficient time for proteolytic digestion by the immobilized pepsin, MALDI-TOF mass spectrometry analysis was performed on the surface of flow cell-2. The resulting peptide masses clearly indicated the presence of IL-1α (Nelson et al., 2000).

The experiments described above indicate that technology is available to couple SPR with mass spectrometry. These methods should be useful for protein-protein interaction mapping. For example, immobilized proteins can be used as hooks for fishing binding partners from complex protein mixtures under native conditions. The coupling of techniques can lead not only to the rapid identification of interacting proteins but will also provide information on the kinetic parameters of the interaction. This approach should serve as an excellent complement to the use of *in vivo* techniques such as the yeast two-hybrid system.

This page intentionally left blank.

Chapter 8
CONCLUSIONS

Proteomics is a rapidly developing field that is heavily focused on technology development. A primary goal has been to provide tools to determine gene function by direct experimentation on proteins. For the past several years, proteomics has been dominated by protein expression mapping via 2D gel electrophoresis and mass spectrometry. New developments in protein fractionation are likely to supplant 2D gel electrophoresis and greatly speed the identification of proteins directly by mass spectrometry. An early example of this is the coupling of liquid chromatography with tandem mass spectroscopy to identify components of the yeast ribosome without the use of 2D gel electrophoresis (Link et al., 1999). In addition, the development of technologies to label proteins within crude lysates in order to quantitate protein levels by mass spectrometry is likely to further advance protein expression mapping. This approach is best illustrated by the ICAT-labeling method which facilities the detection, quantitation and identification of proteins from complex mixtures (Gygi et al., 1999).

Protein-protein interaction mapping has expanded greatly and will be a critical aspect of proteomics studies for the next several years. Knowledge of protein interaction partners provides important information on the possible function of proteins. The protein networks that have been defined to date, notably for yeast, are highly interconnected and include the majority of the proteins encoded by the organism. The fact that nearly all proteins interact with a network of other proteins suggests that our knowledge of gene function will expand exponentially as the function of individual proteins is determined. For instance, if the function of a previously unknown protein is determined experimentally, and that protein interacts with several other proteins in a network, information will be gained on the function of all of these proteins simultaneously.

A fascinating future area of study will be experimental and computational evaluation of the dynamics of protein networks. For example, how do protein complexes and interaction networks change in response to environmental signals or developmental states? How do the networks of

proteins involved in different cellular processes communicate and coordinate activities? Large-scale yeast two-hybrid experiments have indicated many connections between proteins involved in different processes or between proteins from different cellular compartments. These interactions presumably link the functions but further experiments are required to investigate the significance of these connections. A difficult experimental challenge lies in monitoring the dynamics of protein-protein interaction networks within living cells. The use of reagents that evaluate protein interactions and provide an *in vivo* fluorescent signal such as the dihydrofolate reductase protein complementation assay are a step in this direction (Remy and Michnick, 2001).

Membrane proteins continue to challenge proteomics studies on a number of levels. For example, membrane proteins are difficult to resolve by 2D gel electrophoresis and so have posed a problem for protein expression mapping. In addition, membrane proteins are largely excluded from the yeast two-hybrid assay for protein-protein interactions because the interacting proteins must reach the yeast nucleus where they can activate transcription of the reporter genes (Fields and Song, 1989). Membrane proteins have difficulty reaching the nucleus and therefore are underrepresented in the large-scale two-hybrid assays. Surface plasmon resonance is a promising technology for the study of membrane protein interactions. SPR chips containing lipid bilayer surfaces may facilitate the study of membrane protein interactions under native-like conditions. In addition, the coupling of SPR with mass spectrometry may enable the identification of protein interaction partners for membrane proteins from crude protein lysates.

The development of protein chip assays to determine protein function using purified components is a rapidly advancing area. Automated systems for the assay of protein function on chips in parallel for thousands of proteins simultaneously will likely be available in the next few years. These miniaturized arrays will be useful for basic research as well as for diagnostics and drug development. For instance, the combination of protein chips with combinatorial chemistry will allow the simultaneous screening of vast collections of small molecules against vast collections of potential target proteins.

Finally, there is a need to store the enormous amount of data being generated by proteomics experiments. Because of the complexity of analyzing hundreds to thousands of proteins simultaneously, computational methods are needed to visualize and integrate protein expression and protein interaction data. Techniques for visualizing and analyzing protein-protein interaction data have been developed (Schwikowski et al., 2000; Jeong et al., 2001). Further integration of proteomics data with other functional genomics data is required to generate meaningful insights into complex biological processes. For example, it would be useful to computationally integrate

protein-protein interaction data with protein expression data generated by 2D gels as well as mRNA expression data obtained using microarrays (Vidal, 2001). With such integration, one could rapidly assess whether proteins that physically interact are also expressed at similar levels and times within the cell. In addition, as new functional genomics assays are developed, the data obtained will need to be linked to existing data. It is clear that bioinformatics will play an increasing important role in proteomics. Thus, although proteomics has advanced rapidly in the past few years, there remain many experimental and computational challenges for the years ahead.

This page intentionally left blank.

REFERENCES

Abremski, K., Hoess, R., and Sternberg, N. (1983). Studies on the properties of P1 site-specific recombination: evidence for topologically unlinked products following recombination. Cell *32,* 1301-1311.

Ahn, N. G., and Resing, K. A. (2001). Toward the phosphoproteome. Nat. Biotechnol. *19,* 317-318.

Aicher, L., Wahl, D., Arce, A., Grenet, O., and Steiner, S. (1998). New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis *19,* 1998-2003.

Anderson, N. G., and Anderson, N. L. (1996). Twenty years of two-dimensional electrophoresis: past, present and future. Electrophoresis *17,* 443-453.

Anderson, N. L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., and Eacho, P. (1996). The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. *137,* 75-89.

Andersson, L., and Porath, J. (1986). Isolation of phosphoproteins by immobilized metal (Fe^{3+}) affinity chromatography. Anal. Biochem. *154,* 250-254.

Araki, H., Nakanishi, N., Evans, E. R., Matsuzaki, H., Jayaram, M., and Oshima, Y. (1992). Site-specific recombinase, R, encoded by yeast plasmid pSRI. J. Mol. Biol. *225,* 25-37.

Arenkov, P., Kukhtin, A., Gemmell, A., Voloshchuk, S., Chupeeva, V., and Mirzabekov, A. (2000). Protein microchips: Use for immunoassay and enzymatic reactions. Anal. Biochem. *278,* 123-131.

Bai, C., and Elledge, S. J. (1997). Gene identification using the yeast two-hybrid system. Methods Enzymol. *283,* 141-156.

Barabasi, A. L., and Albert, R. (1999). Emergence of scaling in random networks. Science *285,* 509-512.

Barbas, C. F., III, Burton, D. R., Scott, J. K., and Silverman, G. J. (2001). Phage display: A laboratory manual. (Cold Spring Harbor: Cold Spring Harbor Laboratoy Press).

Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. Nature Genet. *12,* 72-77.

Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M., and Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. Proc. Natl. Acad. Sci. USA *92,* 8259-8263.

Bernard, P., and Couturier, M. (1992). Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. J. Mol. Biol. *226,* 735-745.

Berndt, P., Hobohm, U., and Langen, H. (1999). Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. Electrophoresis *20,* 3521-3526.

Brent, R., and Finley, R. L., Jr. (1997). Understanding gene and allele function with two-hybrid methods. Annu. Rev. Genet. *31,* 663-704.

Brown, P. O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat. Genet. *21,* 33-37.

Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc. Natl. Acad. Sci. USA *98,* 7158-7163.

Celia, H., Wilson-Kubalek, E., Milligan, R. A., and Teyton, L. (1999). Structure and function of a membrane-bound murine MHC class I molecule. Proc. Natl. Acad. Sci. USA *96,* 5634-5639.

Cestra, G., Castagnoli, L., Dente, L., Minenkova, O., Petrelli, A., Migone, N., Hoffmuller, U., Schneider-Mergener, J., and Cesareni, G. (1999). The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojananin 1 at distinct sites that display an unconventional binding specificity. J. Biol. Chem. *274,* 32001-32007.

Cochrane, D., Webster, C., Masih, G., and McCafferty, J. (2000). Identification of natural ligands for SH2 domains from a phage display cDNA library. J. Mol. Biol. *297,* 89-97.

Cohen, C., and Parry, D. A. (1994). Alpha-helical coiled coils: more facts and better predictions. Science *264,* 1068.

Cotter, R. J. (1999). The new time-of-flight mass spectrometry. Anal. Chem. *71,* 445A-451A.

Crameri, R., Jaussi, R., Menz, G., and Blaser, K. (1994). Display of expression products of cDNA libraries on phage surfaces. Eur. J. Biochem. 226, 53-58.

Crameri, R., and Suter, M. (1993). Display of biologically active proteins on the surface of filamentous phages: a cDNA cloning system for selection of functional gene products linked to the genetic information responsible for their production. Gene *137,* 69-75.

Cross, F. R. (1995). Starting the cell cycle: what's the point? Curr. Opin. Cell Biol. *6,* 790-797.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23,* 324-328.

Davies, H., Lomas, L., and Austen, B. (1999). Profiling of amyloid b peptide variants using SELDI ProteinChip arrays. BioTechniques *27,* 1258-1261,

de Wildt, R. M. T., Mundy, C. R., Gorick, B. D., and Tomlinson, I. M. (2000). Antibody arrays for high-throughput screening of antibody-antigen interactions. Nat. Biotechnol. *18,* 989994.

DeLano, W. L., Ultsch, M. H., de Vos, A. M., and Wells, J. A. (2000). Convergent solution to binding at a protein-protein interface. Science *287,* 1279-1283.

Dove, S. L., and Hochschild, A. (1998). Conversion of the w subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. Genes Dev. *12,* 745-754.

Ebright, R. H., and Busby, S. (1995). The *E. coli* RNA polymerase a subunit: Structure and function. Curr. Opin. Genet. Dev. *5,* 197-203.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. 0. (2000). Protein function in the post-genome era. Nature *405,* 823-826.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402,* 86-90.

Evans, S. V., and MacKenzie, C. R. (1999). Characterization of protein-glycolipid recognition at the membrane biolayer. J. Molec. Recognit. *12.*

Fayet, O., Ziegelhoffer, T., and Georgopoulos, C. (1989). The groES and groEL heat shock gene products of *Escherichia coli* are essential for bacterial growth at all temperatures. J. Bacteriol. *171,* 1379-1385.

Felici, F., Luzzago, A., Folgori, A., and Cortese, R. (1993). Mimicking of discontinuous epitopes by phage-displayed peptides, II. Selection of clones recognized by a protective monoclonal antibody against the Bordetella pertussis toxin from phage peptide libraries. Gene 128, 21-27.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340,* 245-246.

Fodor, S. P. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science *251,* 767-773.

Folgori, A., Tafi, R., Meola, A., Felici, F., Galfre, G., Cortese, R., Monaci, P., and Nicosia, A. (1994). A general strategy to identify mimotopes of pathological antigens using only random peptide libraries and human sera. EMBO J. *13,* 2236-2243.

Frank, R. (1992). Spot synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. Tetrahedron *48,* 9217-9232.

Fung, E. T., Thulasiraman, V., Weinberger, S. R., and Dalmasso, E. A. (2001). Protein biochips for differential profiling. Curr. Opin. Biotechnol. *12,* 65-69.

Gausepohl, H., Boulin, C., Kraft, M., and Frank, R. W. (1992). Automated multiple peptide synthesis. Pept. Res. *5,* 315-320.

Gauss, C., Kalkum, M., Lowe, M., Lehrach, H., and Klose, J. (1999). Analysis of the mouse proteome. (I) Brain proteins: Separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. Electrophoresis *20,* 575-600.

Gegg, C. V., Bowers, K. E., and Matthews, C. R. (1997). Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. Protein Sci. *6,* 1885-1892.

Geysen, H. M., Meloen, R. H., and Barteling, S. J. (1984). Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. Proc. Natl. Acad. Sci. USA *81,* 3998-4002.

Geysen, H. M., Rodda, S. J., and Mason, T. J. (1986). A *priori* delineation of a peptide which mimics a discontinuous antigenic determinant. Mol. Immunol. *23,* 709-715.

Gmuender, H., Kuratli, K., Di Padova, K., Gray, C. P., Keck, W., and Evers, S. (2001). Gene expression changes triggered by exposure of *Haemophilus influenzae* to novobiocin or ciprofloxacin: combined transcription and translation analysis. Genome Res. *11,* 28-42.

Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000). The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis *21,* 1037-1053.

Grant, P. A., Schieltz, D., Pray-Grant, M. G., Steger, D. J., Reese, J. C., Yates, J. R. r., and Workman, J. L. (1998). A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. Cell *94,* 45-53.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. USA *97,* 9390-9395.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. *17,* 994-999.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. *19,* 1720-1730.

Hartley, J. L., Temple, G. F., and Brasch, M. A. (2000). DNA cloning using in vitro site-specific recombination. Genome Research *10,* 1788-1795.

Hasty, J., and Collins, J. J. (2001). Protein Interactions: Unspinning the web. Nature *411,* 30-31.

Hazbun, T.R., and Fields, S. (2001). Networking proteins in yeast. Proc. Natl. Acad. Sci. USA *98,* 4277-4278.

Hellman, J. D., and Chamberlin, M. J. (1988). Structure and function of bacterial sigma factors. Ann. Rev. Biochem. *57,* 839-872.

Henzel, W. J., Billeci, T. M., Stults, J. T., and Wong, S. C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc. Natl. Acad. Sci. USA *90,* 5011-5015.

Heyman, J. A., Cornthwaite, J., Foncerrada, L., Gilmore, J. R., Gontang, E., Hartman, K. J., Hernandez, C. L., Hood, R., Hull, H. M., Lee, W.-Y., Marcil, R., Marsh, E. J., Mudd, K. M., Patino, M. J., Purcell, T. J., Rowland, J. J., Sindici, M. L., and Hoeffler, J. P. (1999). Genome-scale cloning and expression of individual open reading frames using topoisomerase I-mediated ligation. Genome Res. *9,* 383-392.

Hollinger, P., Prospero, T., and Winter, G. (1993). "Diabodies": Small bivalent and bispecific antibody fragments. Proc. Natl. Acad. Sci. USA *90,* 6444-6448.

Hottiger, M., Gramatikoff, K., Georgiev, O., Chaponnier, C., Schaffner, W., and Hubscher, U. (1995). The large subunit of HIV-1 reverse transcriptase interacts with β-actin. Nucleic Acids Res. *23,* 736-741.

Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F., and Hartl, F. U. (1999). Identification of *in vivo* substrates of the chaperonin GroEL. Nature *402,* 147-154.

Hu, J. C. (2000). A guided tour in protein interaction space: coiled coils from the yeast proteome. Proc. Natl. Acad. Sci. USA *97,* 12935-12936.

Hudson, J. R., Dawson, E. P., Rushing, K. L., Jackson, C. H., Lockshon, D., Conover, D., Lanciault, C., Harris, J. R., Simmons, S. J., Rothstein, R., and Fields, S. (1997). The complete set of predicted genes from *Saccharomyces cerevisiase* in a readily usable form. Genome Research *7,* 1169-1173.

Igarashi, K., Fujita, N., and Ishihama, A. (1991). Identification of a subunit assembly domain in the alpha subunit of *Escherichia coli* RNA polymerase. J. Mol. Biol. *218,* 1-6.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98,* 4569-4574.

Jacobsson, K., and Frykberg, L. (1995). Cloning of ligand-binding domains of bacterial receptors by phage display. BioTechniques *18,* 878-885.

Jacobsson, K., and Frykberg, L. (1996). Phage display shot gun cloning of ligand-binding domains of prokaryotic receptors approaches 100% correct clones. BioTechniques *20,* 1070-1080.

Jacobsson, K., Jonsson, H., Lindmark, H., Guss, B., Lindberg, M., and Frykberg, L. (1997). Shot-gun phage display mapping of two streptococcal cell-surface proteins. Microbiol. Res. *152,* 121-128.

James, P., Halladay, J., and Craig, E. A. (1996). Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. Genetics *144,* 1425-1436.

Janeway, C. A., Jr., and Travers, P. (1994). Immunobiology: The immune system in health and disease. (New York: Garland Publishing).

Jensen, P. K., Pasa-Tolic, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. *71,* 2076-2084.

Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature *411,* 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. Nature *407,* 651-654.

Johnsson, N., and Varshavsky, A. (1994). Split ubiquitin as a sensor of protein interactions in vivo. Proc. Natl. Acad. Sci. USA *91,* 10340-10344.

Jonsson, U., Fagerstam, L., Ivarsson, B., Johnsson, B., Karlsson, R., Lundh, K., Lofas, S., Persson, B., Roos, H., Ronnberg, I., Sjolander, S., Stenberg, E., Stahlberg, R., Urbaniczky, S., Ostlin, H., and Malmqvist, M. (1991). Real-time biospecific interaction analysis using surface plasmon resonance and a sensor chip technology. Biotechniques *11,* 620-627.

Joung, J. K., Ramm, E. I., and Pabo, C. 0. (2000). A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc. Natl. Acad. Sci. USA *97,* 7382-7387.

Katz, B. A. (1997). Structural and mechanistic determinants of affinity and specificity of ligands discovered or engineered by phage display. Annu. Rev. Biophys. Biomol. Struct. *26,* 27-45.

Knoblauch, N. T. M., Rudiger, S., Schonfeld, H. J., Driessen, A. J. M., Schneider-Mergener, J., and Bukau, B. (1999). Substrate specificity of the SecB chaperone. J. Biol. Chem. *274,* 34219-34225.

Lackman, M., Bucci, T., Mann, R. J., Kravets, L. A., Viney, E., Smith, F., Moritz, R. L., Carter, W., Simpson, R. J., Nicola, N. A., Mackwell, K., Nice, E. C., Wilks, A. F., and Boyd, A. W. (1996). Purification of a ligand for the EPH-like receptor HEK using a biosensor-based affinity detection approach. Proc. Natl. Acad. Sci. USA *93,* 2523-2527.

Landy, A. (1989). Dynamic, structural, and regulatory aspects of site-specific recombination. Annu. Rev. Biochem. *58,* 913-949.

Langen, H., Takacs, B., Evers, S., Berndt, P., Lahm, H. W., Wipf, B., Gray, C., and Fountoulakis, M. (2000). Two-dimensional map of the proteome of *Haemophilus infuenzae.* Electrophoresis *21,* 411-429.

Lerner, R. A., Kang, A. S., Bain, J. D., Burton, D. R., and Barbas, C. F. (1992). Antibodies without immunization. Science *258,* 1313-1314.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates III, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. *17,* 676-682.

Liu, Q,, Li, M. Z., Leibham, D., Cortez, D., and Elledge, S. J. (1998). The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. Current Biology *8,* 1300-1309.

Liu, Y., Patricelli, M. P., and Cravatt, B. F. (1999). Activity-based protein profiling: The serine hydrolases. Proc. Natl. Acad. Sci. USA *96,* 14694-14699.

Lueking, A., Horn, M., Eickhoff, H., Bussow, K., Lehrach, H., and Walter, G. (1999). Protein microarrays for gene expression and antibody screening. Anal. Biochem. *270,* 103-111.

Ma, H., Kunes, S., Schatz, P. J., and Botstein, D. (1987). Plasmid construction by homologous recombination in yeast. Gene *58,* 201-216.

MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. Science *289,* 1760-1763.

Mann, M. (1996). A shortcut to interesting human genes: peptide sequence tags, expressed-sequence tags and computers. Trends Biochem. Sci. *21,* 494-495.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285,* 751-753.

Marks, J. D., Hoogenboom, H. R., Bonnert, T. P., McCafferty, J., Griffiths, D., and Winter, G. (1991). By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J. Mol. Biol. *222,* 581-597.

Martzen, M. R., McCraith, S. M., Spinellis, S. L., Torres, F. M., Fields, S., Grayhack, E. J., and Phizicky, E. M. (1999). A biochemical genomics approach for identifying genes by the activity oftheir products. Science *286,* 1153-1155.

Maruyama, I. N., Maruyama, H. I., and Brenner, S. (1994). 1 foo: A I phage vector for the expression of foreign proteins. Proc. Natl. Acad. Sci. USA *91,* 8273-8277.

Mayer, M. L., and Hieter, P. (2000). Protein networks-built by association. Nature Biotechnol. *18,* 1242-1243.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. Proc. Natl. Acad. Sci USA *97,* 4879-4884.

Metcalf, W. W., Jiang, W., and Wanner, B. L. (1994). Use of the rep technique for allele replacement to construct new *Escherichia coli* hosts for maintenance of R6K gamma origin plasmids at different copy numbers. Gene *138,* 1-7.

Mintz, P. J., Patterson, S. D., Neuwald, A. F., Spahr, C. S., and Spector, D. L. (1999). Purification and biochemical characterization of interchromatin granule clusters. EMBO J. *18,* 4308-4320.

Miranker, A. D. (2000). Mass spectrometry of proteins of known mass. Proc. Natl. Acad. Sci. USA *97,* 14025-14027.

Mohler, W. A., and Blau, H. M. (1996). Gene expression and cell fusion analyzed by *lacZ* complementation in mammalian cells. Proc. Natl. Acad. Sci. USA *93,* 12423-12427.

Molina, F., Laune, D., Gougat, C., Pau, B., and Granier, C. (1996). Improved performances of spot multiple peptide synthesis. Pept. Res. *9,* 151-155.

Mooney, J. F., Hunt, A. J., McIntosh, J. R., Liberko, C. A., Walba, D. M., and Rogers, C. T. (1996). Patterning of functional antibodies and other proteins by photolithography of silane monolayers. Proc. Natl. Acad. Sci. USA *93,* 12287-12291.

Morton, T. A., and Myszka, D. G. (1998). Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. Methods Enzymol. *295,* 268-294.

Nelson, R. W., Krone, J. R., and Jansson, O. (1997). Surface plasmon resonance biomolecular interaction analysis mass spectrometry. 1. Chip-based analysis. Anal. Biochem. *69,* 4363-4368.

Nelson, R. W., Nedelkov, D., and Tubbs, K. A. (2000). Biosensor chip mass spectrometry: A chip-based proteomics approach. Electrophoresis *21,* 1155-1163.

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A., and Mann, M. (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. Nat. Genet. *20,* 46-50.

Neubauer, G., and Mann, M. (1999). Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: potentials and limitations. Anal. Chem. *71,* 235-242.

Newman, J. R. S.•(2000). A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae.* Proc. Natl. Acad. Sci. USA *97,* 13203-13208.

Nilsson, M., Frykberg, L., Flock, J. I., Pei, L., Lindberg, M., and Guss, B. (1998). A fibrinogen-binding protein of *Staphylococcus epidermidis.* Infect. & Immun. *66,* 2666-2673.

Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. Proc. Natl. Acad. Sci. USA *96,* 6591-6596.

Oda, Y., Nagasu, T., and Chait, T. (2001). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. Nat. Biotechnol. *19,* 379-382.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250,* 4007-4021.

Ostergaard, M., Wolf, H., Orntoft, T. F., and Celis, J. E. (1999). Psoriasin (S100A7): A putative urinary marker for the follow-up of patients with bladder squamous cell carcinomas. Electrophoresis *20,* 349-354.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA *96,* 2896-2901.

Padlan, E. A. (1993). Anatomy of the antibody molecule. Mol. Immunol. *31,* 169-217.

Page, M. J., Amess, B., Townsend, R. R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M. J., Stein, R. C., Waterfield, M. D., Davies, S. C., and O'Hare, M. J. (1999). Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. Proc. Natl. Acad. Sci. USA *96,* 12589-12594.

Palzkill, T., Huang, W., and Weinstock, G. M. (1998). Mapping protein-ligand interactions using whole genome phage display libraries. Gene *221,* 79-83.

Pandey, A., and Mann, M. (2000). Proteomics to study genes and genomes. Nature *405,* 837-846.

Pasqualini, R., Koivunen, E., and Ruoslahti, E. (1997). Alpha V integrins as receptors for tumor targeting by circulating ligands. Nat. Biotechnol. *15,* 542-546.

Pasqualini, R., and Ruoslahti, E. (1996). Organ targeting in vivo using phage display peptide libraries. Nature *380,* 364-366.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252,* 809-817.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA *96,* 4285-4288.

Pelletier, J. N., Campbell-Valois, F.-X., and Michnick, S. W. (1998). Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. Proc. Natl. Acad. Sci. USA *95,* 12141-12146.

Peltier, J. B., Friso, G., Kalume, D. E., Roepstorff, P., Nilsson, F., Adamska, I., and van Wijk, K. J. (2000). Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. Plant Cell *12,* 303-304.

Persson, M. A., Caothien, R. H., and Burton, D. R. (1991). Generation of diverse high-affinity human monoclonal antibodies by repertoire cloning. Proc. Natl. Acad. Sci. USA *88,* 2432-2436.

Petersen, G., Song, D., Hugle-Dorr, B., Oldenburg, I., and Bautz, E. K. F. (1995). Mapping of linear epitopes recognized by monoclonal antibodies with gene-fragment pahge display libraries. Mol. Gen. Genet. *249,* 425-431.

Piddock, L. J. V., Walters, R. N., and Diver, J. M. (1990). Correlation of quinolone MIC and inhibition of DNA, RNA and protein synthesis and induction of the SOS response in *Escherichia coli.* Antimicrob. Agents Chemother. *34,* 2331-2336.

Ptashne, M. (1992). A Genetic Switch, 2 Edition (Cambridge, Mass.: Blackwell Scientific Publications).

Pullen, S. S., Dang, T. T. A., Crute, J. J., and Kehry, M. R. (1999). CD40 signalling through tumor necrosis factor receptor-associated factors (TRAFs). J. Biol. Chem. *274,* 14246-14254.

Rabilloud, T., Valette, C., and Lawrence, J. J. (1994). Two-dimensional electrophoresis of basic proteins with equilibrium isoelectric focusing in carrier ampholyte-pH gradients. Electrophoresis *15,* 1552-1558.

Rader, C., and Barbas, C. F., III (1997). Phage display of combinatorial antibody libraries. Curr. Opin. Biotechnol. 8, 503-508.

Rain, J.-C., Selig, L., DeReuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of *Helicobacter pylori.* Nature *409,* 211-215.

Rajotte, D., Arap, W., Hagedorn, M., Koivunen, E., Pasqualini, R., and Ruoslahti, E. (1998). Molecular heterogeneity of the vascular endothelium revealed by in vivo phage display. J. Clin. Invest. *102,* 430-437.

Rasched, I., and Oberer, E. (1986). Ff coliphages: Structural and functional relationships. Microbiol. Rev. *50,* 401-427.

Rasmussen, H. H., Orntoft, T. F., Wolf, H., and Celis, J. E. (1996). Towards a comprehensive database of proteins from the urine of patients with bladder cancer. J. Urol. *155,* 2113-2119.

Reineke, U., Kramer, A., and Schneider-Mergener, J. (1999). Antigen sequence- and library-based mapping of linear and discontinuous protein-protein interaction sites by spot synthesis. Curr. Top. Microbiol. Immunol. *243,* 23-36.

Reineke, U., Sabat, R., Misselwitz, R., Welfle, H., Volk, H.-D., and Schneider-Mergener, J. (1999). A synthetic mimic of a discontinuous binding site on interleukin-10. Nat. Biotechnol. *17,* 271-275.

Reineke, U., Volkmer-Engert, R., and Schneider-Mergener, J. (2001). Applications of peptide arrays prepared by the SPOT-technology. Curr. Opin. Biotechnol. *12,* 59-64.

Remy, I., and Michnick, S. W. (1999). Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. Proc. Natl. Acad. Sci. USA *96,* 5394-5399.

Remy, I., and Michnick, S. W. (2001). Visualization of biochemical networks in living cells. Proc. Natl. Acad. Sci. USA *98,* 7678-7683.

Ren, Z. J., Lewis, G. K., Wingfield, P. T., Locke, E. G., Steven, A. C., and Black, L. W. (1996). Phage display of intact domains at high copy number: a system based on SOC, the smaller outer capsid protein of bacteriophage T4. Protein Sci. *5,* 1833-1843.

Rich, R. L., and Myszka, D. G. (2000). Advances in surface plasmon resonance biosensor analysis. Curr. Opinion Biotechnol. *11,* 54-61.

Riechmann, L., and Holliger, P. (1997). The C-terminal domain of TolA is the coreceptor for filamentous phage infection of *E. coli.* Cell *90,* 351-360.

Roos, M., Soskic, V., Poznanovic, S., and Godovac-Zimmermann, J. (1998). Post-translational modifications of endothelin receptor B from bovine lungs analyzed by mass spectrometry. J. Biol. Chem. *273,* 924-931.

Rosenberg, A., Griffin, G., Studier, F. W., McCormick, M., Berg, J., and Mierendorf, R. (1996). T7Select phage display system: a powerful new protein display system based on bacteriophage T7. inNovations 6, 1-6.

Rossi, F., Charlton, C. A., and Blau, H. M. (1997). Monitoring protein-protein interactions in intact eukaryotic cells by 0-galactosidase complementation. Proc. Natl. Acad. Sci. USA *94,* 8405-8410.

Rossi, F. M. V., Blakely, B. T., and Blau, H. M. (2000). Interaction blues: protein interactions monitored in live mammalian cells by fl-galactosidase complementation. Trends Cell Biol. *10,* 119-122.

Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager K, Gerstein, M., Miller, P., Roeder, G. S., and Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature *402,* 362-363.

Rout, M. P., Aitchinson, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. *148,* 635-651.

Rubanyi, G. M., and Polokoff, M. A. (1994). Endothelins: molecular biology, biochemistry, physiology, and pathophysiology. Pharmacol. Rev. *46,* 325-415.

Russell, M. (1991). Filamentous phage assembly. Mol. Microbiol. *5,* 1607-1613.

Santini, C., Brennan, D., Mennuni, C., Hoess, R. H., Nicosia, A., Cortese, R., and Luzzago, A. (1998). Efficient display of an HCV cDNA expression library as C-terminal fusion to the capsid protein D of bacteriophage lambda. J. Mol. Biol. *282,* 125-135.

Santoni, V., Molloy, M., and Rabilloud, T. (2000). Membrane proteins and proteomics: Un amour impossible? Electrophoresis *21,* 1054-1070.

Sblattero, D., and Bradbury, A. (2000). Exploiting recombination in single bacteria to make large phage antibody libraries. Nat. Biotechnol. *18,* 75-80.

Sche, P. P., McKenzie, K. M., White, J. D., and Austin, D. J. (1999). Display cloning: functional identification of natural product receptors using cDNA-phage display. Chem. Biol. *6,* 707-716.

Schuck, P. (1997). Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules. Annu. Rev. Biophys. Biomol. Struct. *26,* 541-566.

Schultz, J., Hoffmuller, U., Krause, G., Ashurst, J., Macias, M. J., Schmieder, P., Schneider-Mergener, J., and Oschknat, H. (1998). Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. Nat. Struct. Biol. *5,* 19-24.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. Nat. Biotechnol. *18,* 1257-1261.

Selkoe, D. J. (1998). The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. Trends Cell Biol. *8,* 447-453.

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. *6,* 639-645.

Shuman, S. (1992a). DNA strand transfer reactions catalyzed by vaccinia topoisomerase I. J. Biol. Chem. *267,* 8620-8627.

Shuman, S. (1994). Novel approach to molecular cloning and polynucleotide synthesis using vaccinai DNA topoisomerase. J. Biol. Chem. *269,* 32678-32684.

Shuman, S. (1992b). Two classes of DNA end-joining reactions catalyzed by vaccinia topoisomerase I. J. Biol. Chem. *267,* 16755-16758.

Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. (2000). Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Reports *1,* 287-292.

Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228,* 1315-1317.

Smith, G. P., and Petrenko, V. A. (1997). Phage display. Chem. Rev. *97,* 391-410.

Soskic, V., Gorlach, M., Poznanovic, S., Boehmer, F. D., and Godovac-Zimmermann, J. (1999). Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor b receptor. Biochemistry *38,* 1757-1764.

Steinberg, T. H., Haugland, R. P., and Singer, V. L. (1996). Applications of SYPRO orange and SYPRO red protein gel stains. Anal. Biochem. *239,* 238-245.

Sternberg, N., Hamilton, D., Austin, S., Yarmolinsky, M., and Hoes, R. (1981). Site-specific recombination and its role in the life cycle of P1. Cold Spring Harbor Symp. Quant. Biol. *45,* 297-309.

Surdo, P. L., Bottomley, M. J., Arcaro, A., Siegal, G., Panayotou, G., Sankar, A., Gaffney, P. R. J., Riley, A. M., Potter, B. V. L., Waterfield, M. D., and Driscoll, P. C. (1999). Structural and biochemical evaluation of the interaction of the phosphatidylinositol 3-kinase p85 Src homology 2 domains with phosphoinositides and inositol polyphosphates. J. Biol. Chem. *274,* 15678-15685.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., ., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Venter, J. C., and al., e. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature *388,* 539-547.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae.* Nature *403,* 623-627.

Vaughan, T. J., Williams, A. J., Pritchard, K., Osbourn, J. K., Pope, A. R., Earnshaw, J. C., McCafferty, J., Hodits, R. A., Wilton, J., and Johnson, K. S. (1996). Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library. Nat. Biotechnol. *14,* 309-314.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, J., D.E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptosome. Cell *88,* 243-251.

Vidal, M. (2001). A biological atlas of functional maps. Cell *104,* 333-339.

Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science *287,* 116-122.

Wallin, E., and Von Heijne, G. (1998). Prot. Sci. *1998,* 1029-1038.

Wigge, P. A., Jensen, O. N., Holmes, S., Soues, S., Mann, M., and Kilmartin, J. V. (1998). Analysis of the Saccharomyces spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. J. Cell Biol. *141,* 967-977.

Williams, C., and Addona, T. A. (2000). The integration of SPR biosensors with mass spectrometry: possible applications for proteome analysis. Trends Biotechnol. *18,* 45-48.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature *379,* 466-469.

Winter, D., Podtelejnikiov, A. V., Mann, M., and Li, R. (1997). The complex containing actin-related proteins Arp2 and Arp3 is required for the motility and integrity of yeast actin patches. Curr. Biol. *7,* 519-529.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., and Davis, R. W., et al. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science *Science,* 901-906.

Wolf, E., Kim, P. S., and Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci. *6,* 1179-1189.

Wright, G. L., Cazares, L. H., Leung, S.-M., Nasim, S., Adam, B.-L., Yip, T.-T., Schellhammer, P. F., Gong, L., and Vlahou, A. (2000). ProteinChip surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. Prostate Cancer and Prostatic Diseases *2,* 264-276.

Yao, Z.-J., Kao, M. C. C., and Chung, M. C. M. (1995). Epitope identification by polyclonal antibody from phage-displayed random peptide library. J. Protein Chem. *14,* 161-166.

Yates III, J. R. (2000). Mass spectrometry from genomics to proteomics. Trends in Genet. *16,* 5-8.

Young, R. A., and Davis, R. W. (1983). Efficient isolation of genes using by using antibody probes. Proc. Natl. Acad. Sci USA *80,* 1194-1198.

Zhou, H., Watts, J. D., and Aebersold, R. (2001). A systematic approach to the analysis of protein phosphorylation. Nat. Biotechnol. *19,* 375-378.

Zozulya, S., Lioubin, M., Hill, R. J., Abram, C., and Gishizky, M. L. (1999). Mapping signal transduction pathways by phage display. Nat. Biotechnol. *17,* 1193-1198.

Zucchi, I., Bini, L., Valaperta, R., Ginestra, A., Albani, D., Susani, L., Sanchez, J. C., Liberatori, S., Magi, B., Raggiaschi, R., Hochstrasser, D. F., Pallini, V., Vezzoni, P., and Dulbecco, R. (2001). Proteomic dissection of dome formation in a mammary cell line: Role of tropomyosin-5b and maspin. Proc. Natl. Acad. Sci. USA *98,* 5608-5613.

INDEX